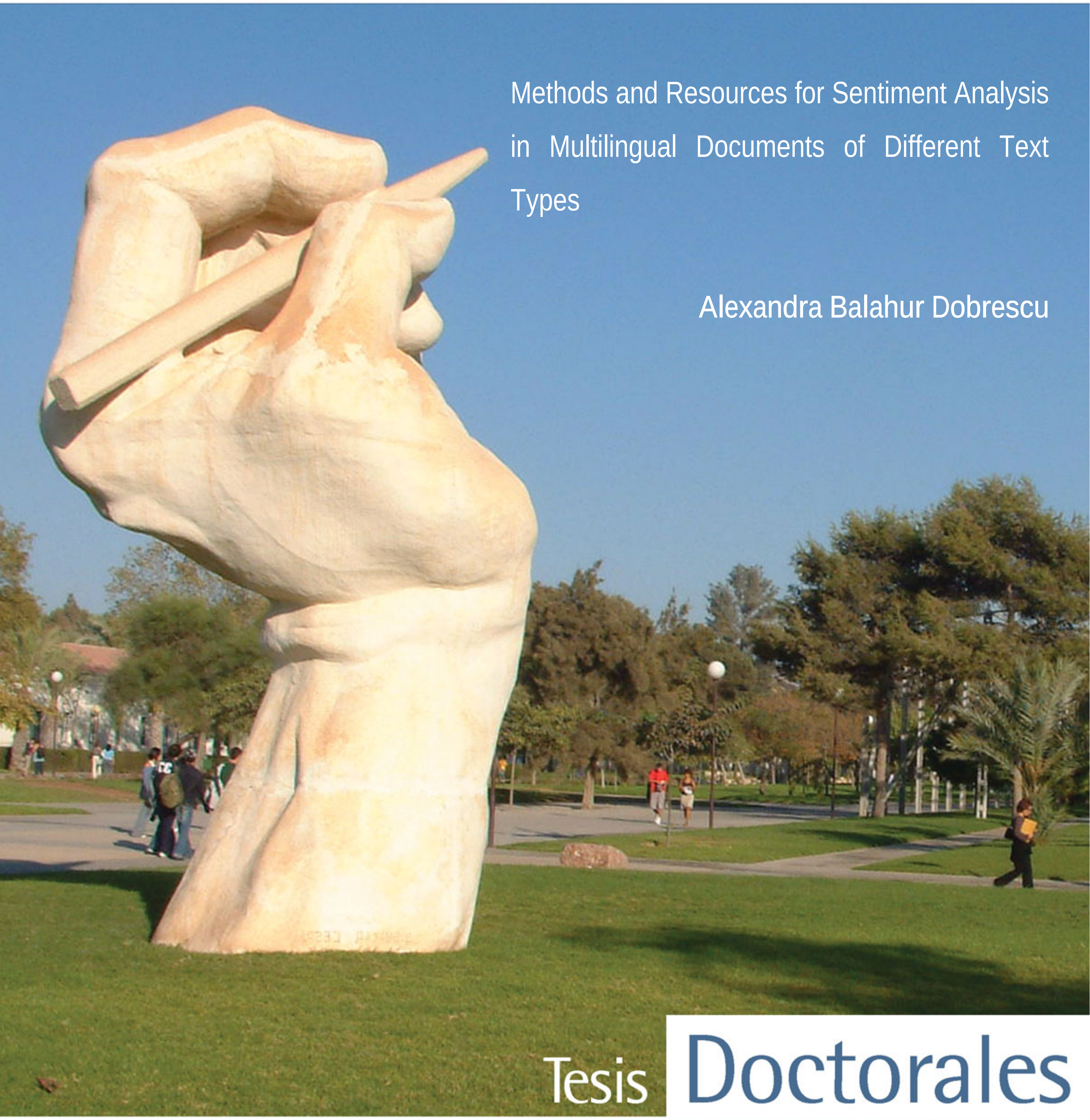


Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types

Alexandra Balahur Dobrescu

Universitat d'Alacant
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos

Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types

PhD Thesis

Alexandra Balahur Dobrescu
2011

Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types

Dissertation

Presented to the Department of Software and Computing Systems, University of Alicante, in partial fulfillment of the requirements for the title of

Doctor of Philosophy

Author: **Alexandra Balahur Dobrescu**

Supervisors: Prof. Dr. **Andrés Montoyo**

Dr. **Ralf Steinberger**

Imagen Portada "EL GRITO" (fragmento), Eduard Munch (1893)

“Whereas our argument shows that the power and capacity of learning exists in the soul already, and that just as the eye was unable to turn from darkness to light without the whole body, so too the instrument of knowledge can only by the movement of the whole soul be turned from the world of becoming into that of being, and learn by degrees to endure the sight of being and of the brightest and best of being, or in other words, of the good.”

Plato, The Republic, Book 7, section 7 (The Myth of the Cave)

ACKNOWLEDGEMENTS

First of all, I would like to thank my main advisor Prof. Dr. Andrés Montoyo, for all the knowledge and advice he has shared with me along these almost 4 years. I would like to thank him, Prof. Dr. Manuel Palomar and Prof. Dr. Patricio Martínez-Barco for all the support they have given me in the past years. I would like to thank the University of Alicante for the PhD scholarship and all the people at the Department of Software and Computing Systems, in the Natural Language Processing Group, for their support and friendship. I would especially like to thank Ester Boldrini and Elena Lloret, with whom I had a fruitful collaboration on numerous occasions.

I would like to thank Dr. Ralf Steinberger, my co-advisor, for all the knowledge and experience he has shared with me, as well as all the advice he has given me during the past 2 years. Together with him, I would like to thank the entire OPTIMA team at the European Commission’s Joint Research Centre, led by Erik van der Goot, for everything they taught me during my traineeship in Ispra (Italy), for their valuable advice, their friendship and the collaboration in the development of various resources and methods for sentiment analysis. I would especially like to thank Dr. Mijail Kabadjov and Dr. Josef Steinberger, with whom I have collaborated on various research papers in opinion summarization, during my traineeship and after.

I would like to thank Prof. Dr. Dietrich Klakow for receiving me as an intern at the Spoken Language Systems Department at the University of Saarbrücken (Germany) and for all the knowledge he and his group shared with me during my traineeship there. I would like to thank Dr. Michael Wiegand for his collaboration on various sentiment analysis projects.

I would like to thank my wonderful family – my mother Doina, my father Paul and my sister Elvira for raising me to be curious, ambitious and perseverant, to enjoy science and challenges, to never be afraid to ask, never be scared to look for answers and always aim for “excelsior”. I would like to thank all my family and friends, here, at home and in the world for their support and their advice.

I would like to thank Jesús Hermida, a very special person in my life, for his support, advice and the knowledge he has shared with me.

Finally, I would like to thank all the people that I have come across at different events in the past years, who with their suggestions and critiques helped to improve my research and/or life approach, as well as those who believed in me or doubted me, making me wiser and stronger.

AGRADECIMIENTOS

En primer lugar, quisiera agradecer a mi tutor principal el Prof. Dr. Andrés Montoyo, por todos los conocimientos y consejos que ha compartido conmigo a lo largo de estos casi cuatro años. Me gustaría darle las gracias a él y a Prof. Dr. Manuel Palomar y Prof. Dr. Patricio Martínez, por todo el apoyo que me han dado en los últimos años. Gracias a la Universidad de Alicante por la beca de doctorado y a toda la gente en el Departamento de Lenguajes y Sistemas Informáticos, y en especial al Grupo de Procesamiento del Lenguaje Natural, por su apoyo y amistad. Me gustaría dar las gracias a Ester Boldrini y Elena Lloret, con quien tuve una fructífera colaboración en numerosas ocasiones.

Me gustaría agradecer al Dr. Ralf Steinberger, mi co-tutor, por todo el conocimiento y la experiencia que ha compartido conmigo, así como todos los consejos que me ha dado durante los últimos 2 años. Junto con él, me gustaría dar las gracias a todo el equipo de OPTIMA del Joint Research Centre de la Comisión Europea, dirigido por el Erik van der Goot, por todo lo que me enseñó durante mi periodo de prácticas en Ispra (Italia), por sus valiosos consejos, su amistad y la colaboración en el desarrollo de los diversos recursos y métodos para el análisis de sentimientos. En especial me gustaría dar las gracias a Dr. Mijail Kabadjov y Dr. Josef Steinberger, con los que he colaborado en diversas investigaciones sobre las técnicas de resúmenes de opinión, durante mi periodo de prácticas y después.

Me gustaría agradecer al Prof. Dr. Dietrich Klakow por darme la oportunidad de hacer una estancia de investigación en el Departamento de Sistemas de Habla (en inglés "Spoken Language Systems") de la Universidad de Saarbrücken (Alemania), y por todo el conocimiento que su grupo y él mismo compartieron conmigo durante mi periodo de prácticas. Me gustaría dar las gracias también a Dr. Michael Wiegand por su colaboración en varios proyectos de análisis de emociones, durante mi estancia en Saarbrücken y después.

Me gustaría dar gracias a mi maravillosa familia – mi madre Doina, mi padre Paul y mi hermana Elvira por apoyar mi curiosidad, ambición y perseverancia, por enseñarme a disfrutar de la ciencia y los desafíos, a atreverme a preguntar, no tener miedo a buscar respuestas y siempre tratar de buscar la excelencia en lo que hago, llegar a un "excelsior". Me gustaría dar las gracias a toda mi familia y amigos, aquí, en casa y en el mundo por su apoyo y sus consejos.

Me gustaría dar las gracias a Jesús Hermida, una persona muy especial en mi vida, por su apoyo, el asesoramiento y el conocimiento que ha compartido conmigo.

Por último, me gustaría dar las gracias a toda la gente que me he encontrado en diferentes eventos en los últimos años, que con sus sugerencias y críticas ayudaron a mejorar mi investigación y/o enfoque de la vida. Asimismo, gracias a todos aquellos que creyeron en mí o dudaron de mí, ayudándome así a ser más sabia y más fuerte.

ABSTRACT

The present doctoral thesis deals with the issues and challenges involved in the development of methods and resources for the Natural Language Processing (NLP) task of sentiment analysis.

Specifically, the first aim is to develop adequate techniques for the automatic detection and classification of directly, indirectly or implicitly-expressed sentiment in texts of different types (reviews, newspaper articles, dialogues/debates and blogs), in different languages. The second aim is to apply the sentiment analysis methods proposed in the context or jointly with other NLP tasks and propose adequate techniques to tackle the issues raised by the peculiarities of affect expression in these tasks.

In order to achieve the proposed objectives, the work presented has been structured around answering five research questions. Following is a description of the questions and a summary of the answers we have given in the present thesis.

1. *How can sentiment analysis and, in a broader perspective, opinion mining be defined in a correct way? What are the main concepts to be treated in order to create a good definition that can be used to appropriately define the task and subsequently propose correct methods to tackle it?*

In Chapter 2, we define the main concepts we will frequently employ throughout this thesis. We first present an overview of the definitions given in the NLP literature to the related tasks of subjectivity analysis, sentiment analysis, opinion mining, appraisal/attitude analysis, and emotion detection. We subsequently present the definitions of the terms that are related to these tasks, both in well-established dictionaries, as well as the research literature in the field. Finally, we propose an operational definition that is consistent with the manner in which the different terms related to sentiment analysis are defined. In Chapter 3, we present the state of the art in the field and show that depending on the final aim of the application, the tasks involving sentiment analysis are defined and tackled in a different manner.

The subsequent research questions we address in this thesis are:

2. *Can sentiment analysis be performed using the same methods, for all text types? What are the peculiarities of the different text types and how do they influence the methods to be used to tackle it? Do we need special resources for different text types?*

3. *Can the same language resources be used in other languages (through translation)? How can resources be extended to other languages?*

In Chapter 4, we present the peculiarities of different text types (reviews, newspaper articles, blogs, political debates), analyze them and propose adequate techniques to address them at the time of performing sentiment analysis. In the cases where no generally-accepted definition of the sentiment analysis task exists for a specific textual genre, we propose new definitions and annotate new resources accordingly. We present different methods and resources we built for the task of sentiment analysis in different text types, in different languages (English, Spanish, German). In each of the genres studied, we evaluate our approaches correspondingly, both in-house, as well as in international competitions. We show that the techniques employed are robust enough to obtain good results, even in the case where the original texts are in a language for which we do not have any resources available and for the treatment of which we employ translation engines. Finally, given the results obtained, we show that our approaches perform at the level of state-of-the-art systems and in many cases outperform them.

4. *How can we deal with opinion in the context of traditional tasks? How can we adapt traditional tasks (Information Retrieval, Question Answering, Text Summarization) in the context of opinionated content? What are the "new" challenges in this context?*

In Chapter 4, we only concentrate on the task of sentiment analysis as a standalone challenge, omitting the steps required in order to obtain the texts on which the sentiment analysis methods were applied or eliminating redundancy in the information obtained. However, in a real-world application scenario, automatically detecting the opinion expressed in a text is often not the first, neither the last task to be performed. In order to analyze the sentiment found in different texts, the documents must firstly be retrieved. Additionally, the results of the automatic sentiment analysis may still contain a high volume of information, with much redundancy. Bearing in mind these necessities, in Chapter 5 we study methods to combine opinion mining with question answering and summarization. We show that performing traditional tasks in the context of opinionated text has many challenges and that systems that were designed to work exclusively with factual data are not able to cope with opinion questions. Thus, we propose new methods and techniques to adapt question answering and summarization systems to deal with opinionated content. Additionally, we create and annotate appropriate resources for the evaluation of the proposed methods. Finally, we evaluate our approaches, as well as the impact of using different tools and resources in these tasks. Our evaluations, both in in-house experiments, as well as through the participation in international competitions, show that the proposed methodologies are appropriate for tackling question answering and summarization in the context of opinionated texts.

The last research question we address in this thesis is:

5. *Can we propose a model to detect emotion (as a component of sentiment) from text, in the cases where it is expressed implicitly, requiring world knowledge for its detection?*

As we will see throughout this thesis, sentiments can be explicitly or implicitly present in texts. While in the first case, lexical clues may be found in the text indicating the presence of sentiment, through sentiment-bearing words, in the second case, the emotion underlying the sentiment is not explicitly stated through the use of affective words. In these situations, the emotion is only inferable based on commonsense knowledge (i.e. emotion is not explicitly, but implicitly expressed by the author, by presenting situations which most people, based on commonsense knowledge, associate with an emotion, like "going to a party", "seeing your child taking his/her first step" etc.). Motivated by the fact that most work in sentiment analysis has been done only in view of the existence of lexical clues for sentiment detection and classification, and having seen the limitations of such models, in Chapter 6 of the thesis, we present our contribution to the issue of automatically detecting emotion expressed in text in an implicit manner. The initial approach is based on the idea that emotion is triggered by specific concepts, according to their *relevance*, seen in relation to the basic needs and motivations, underpinning our idea on the Relevance Theory. The second approach we propose is based on the Appraisal Theory models. The general idea behind it is that emotions are most of the times not explicitly stated in texts, but results from the interpretation (appraisal) of the actions contained in the situation described, as well as the properties of their actors and objects. Thus, we set up a framework for representing situations described in text as chains of actions (with their corresponding actors and objects), and their corresponding properties (including the affective ones), according to commonsense knowledge. We show the manner in which the so-called "appraisal criteria" can be automatically detected from text and how additional knowledge on the properties of the concepts involved in such situations can be imported from commonsense knowledge bases. Finally, we demonstrate through an extensive evaluation that such a representation is useful to obtain an accurate label of the emotion expressed in text, without any linguistic clue being present therein, increasing the recall of systems performing sentiment analysis from texts.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

Motto: *"Human behavior flows from three main sources: desire, emotion and knowledge." (Plato)*

1.1. BACKGROUND

The era in which we live has been given many names. "Global village", "technotronic era", "post-industrial society", "information society", "information age", and "knowledge society" are just a few of the terms that have been used in an attempt to describe the deep changes that have occurred in the lives of societies and people worldwide as a result of the fast development of ICT technologies, the access to Internet and its transformation into a Social Web. In this new context, having access to large quantities of information is no longer an issue, as there are terabytes of new information produced on the Web every day that are available to any individual with an Internet connection. In contrast to older times, when finding sources of information was the key problem to companies and individuals, today's information society challenges companies and individuals to *create and employ mechanisms to search and retrieve* ***relevant*** *data from* the huge quantity of available information and *mine* it to *transform* it into ***knowledge***, which they can use to their advantage. As opposed to the past, when this advantage was a question of finding sources of information, in today's society, which is flooded by data that is changing at a rapid pace, the advantage is given by the ***quality*** (accuracy, reliability) of the extracted knowledge and its ***timeliness***. For the era in which we live, information has become the main trading object. In this context, having at hand high quality and timely information is crucial to all the spheres of human activity: social, political, and economic, to name just a few.

However, in many cases, the relevant information is not found in structured sources (i.e. tables or databases), but in unstructured documents, written in human language. The high quantity of such data requires the use of automatic processing techniques. The discipline that deals with the automatic treatment of natural language in text or speech is called Natural Language Processing (NLP). NLP is part of the research area of Artificial Intelligence (AI), which is defined as "the science and engineering of making intelligent machines" (McCarthy, 1959), by simulating the mechanisms of human intelligence. The goal of Artificial Intelligence, as it was stated in the 1950s, is to create machines that are capable of passing the Turing Test. "Roughly speaking, a computer will have passed the Turing Test if it can engage in conversations indistinguishable from that of a human's" (Lee, 2004). In order to achieve this goal, NLP deals with the text analysis at different levels: *phonologic* (sounds), *lexical* (words), *morphologic* (parts of speech), *syntactic* (representation of the structure of the sequence of lexical units based on their dependency), *semantic* (logical structure representing the meaning expressed) and *pragmatic* (studying the influence of the context and the world knowledge on the general meaning of the text). NLP contains many research areas. Each of these constitute either general NLP problems, which need to be solved in any application areas (Word Sense Disambiguation, Co-reference resolution), or that have been set up in the view of a specific end application (Information Retrieval, Information Extraction, Question Answering, Text Summarization, Machine Translation).

Traditionally, the application areas of NLP were designed for the treatment of factual (exact) data. Nowadays, however, factual information is no longer the main source from which crucial knowledge is extracted.

The present is marked by the growing influence of the Social Web (the web of interaction and communication) on the lives of people worldwide. More than ever before, people are more than willing and happy to share their lives, knowledge, experience and thoughts with the entire world, through blogs, forums, wikis, review sites or microblogs. They are actively participating to events, by expressing their opinions on them, by commenting on the news appearing and the events that take place in all spheres of the society. The large volume of subjective information present on the Internet, in reviews, forums, blogs, microblogs and social network communications has produced an important shift in the manner in which people communicate, share knowledge and emotions and influence the social, political and economic behavior worldwide. In consequence, this new reality has led to important transformations in the manner, extent and rapidness in which news and their associated opinions circulate, leading to new and challenging social, economical and psychological phenomena.

In order to study these phenomena and address the issue of extracting the crucial knowledge that nowadays is contained in opinionated data, new fields of research were born in Natural Language Processing (NLP), aiming at detecting subjectivity in text and/or extracting and classifying opinions into different sets (usually positive, negative and neutral). The main tasks that were tackled in NLP are **subjectivity analysis** (dealing with "private states" (Banfield, 1982), a term that encloses sentiment, opinions, emotions, evaluations, beliefs and speculations) **sentiment analysis** and **opinion mining**, although different terminologies have been used to denote the approaches taken (e.g. review mining, appraisal extraction) and sentiment analysis and opinion mining have been used interchangeably, as they are considered by some authors to point to the same task (Pang and Lee, 2008). A closely related task is also **emotion detection,** dealing with the classification of texts according to the emotion expressed. All these research areas are part of the wider field in Artificial Intelligence denominated **affective computing** (Picard, 1995).

This thesis deals with the task of sentiment analysis, in the context of multilingual documents of different text types. Specifically, the work we will present throughout the following chapters concentrates on answering the following research questions:

1. *How can sentiment analysis and, in a broader perspective, opinion mining be defined in a correct way? What are the main concepts to be treated in order to create a good definition that can be used to appropriately delimitate the task and subsequently propose correct methods to tackle it?*

In Chapter 2, we define the main concepts we will frequently employ throughout this thesis. We first present an overview of the definitions given in the NLP literature to the related tasks of subjectivity analysis, sentiment analysis, opinion mining, appraisal/attitude analysis, and emotion detection. We subsequently present the definitions of the terms that are related to these tasks, both in well-established dictionaries, as well as the research literature in the field. Finally, we propose an operational definition that is consistent with the manner in which the different terms related to sentiment analysis are defined. In Chapter 3, we present the state of the art in the field and show that depending on the final aim of the application, the tasks involving sentiment analysis are defined and tackled in a different manner.

The subsequent research questions we address in this thesis are:

2. *Can sentiment analysis be performed using the same methods, for all text types? What are the peculiarities of the different text types and how do they influence the methods to be used to tackle it? Do we need special resources for different text types?*
3. *Can the same language resources be used in other languages (through translation)? How can resources be extended to other languages?*

In Chapter 4, we present the peculiarities of different text types (reviews, newspaper articles, blogs, political debates), analyze them and propose adequate techniques to address them at the time of performing sentiment analysis. In the cases where no generally-accepted definition of the sentiment analysis task exists for a specific textual genre, we propose new definitions and annotate new resources accordingly. We present different methods and resources we built for the task of sentiment analysis in different text types, in different languages (English, Spanish, German). In each of the genres studied, we evaluate our approaches correspondingly, both in-house, as well as in international competitions. We show that the techniques employed are robust enough to obtain good results, even in the case where the original texts are in a language for which we do not have any resources available and for the treatment of which we employ translation engines. Finally, given the results obtained, we show that our approaches perform at the level of state-of-the-art systems and in many cases outperform them.

4. *How can we deal with opinion in the context of traditional tasks? How can we adapt traditional tasks (Information Retrieval, Question Answering, Text Summarization) in the context of opinionated content? What are the "new" challenges in this context?*

In Chapter 4, we only concentrate on the task of sentiment analysis as a standalone challenge, omitting the steps required in order to obtain the texts on which the sentiment analysis methods were applied or eliminating redundancy in the information obtained. However, in a real-world application scenario, automatically detecting the opinion expressed in a text is often not the first, neither the last task to be performed. In order to analyze the sentiment found in different texts, the documents must firstly be retrieved. Additionally, the results of the automatic sentiment analysis may still contain a high volume of information, with much redundancy. Bearing in mind these necessities, in Chapter 5 we study methods to combine opinion mining with question answering and summarization. We show that performing traditional tasks in the context of opinionated text has many challenges and that systems that were designed to work exclusively with factual data are not able to cope with opinion questions. Thus, we propose new methods and techniques to adapt question answering and summarization systems to deal with opinionated content. Additionally, we create and annotate appropriate resources for the evaluation of the proposed methods. Finally, we evaluate our approaches, as well as the impact of using different tools and resources in these tasks. Our evaluations, both in in-house experiments, as well as through the participation in international competitions, show that the proposed methodologies are appropriate for tackling question answering and summarization in the context of opinionated texts.

The last research question we address in this thesis is:

5. *Can we propose a model to detect emotion (as a component of sentiment) from text, in the cases where it is expressed implicitly, requiring world knowledge for its detection?*

As we will see throughout this thesis, sentiments can be explicitly or implicitly present in texts. While in the first case, lexical clues may be found in the text indicating the presence of sentiment, through sentiment-bearing words, in the second case, the emotion underlying the sentiment is not explicitly stated through the use of affective words. In these situations, the emotion is only inferable based on commonsense knowledge (i.e. emotion is not explicitly, but implicitly expressed by the author, by presenting situations which most people, based on commonsense knowledge, associate with an emotion, like “going to a party”, “seeing your child taking his/her first step” etc.). Motivated by the fact that most work in sentiment analysis has been done only in view of the existence of lexical clues for sentiment detection and classification, and having seen the limitations of such models, in Chapter 6 of the thesis, we present our contribution to the issue of automatically detecting emotion expressed in text in an implicit manner. The initial approach is based on the idea that emotion is triggered by specific concepts, according to their *relevance*, seen in relation to the basic needs and motivations, underpinning our idea on the Relevance Theory. The second approach we propose is based on the Appraisal Theory models. The general idea behind it is that emotions are most of the times not explicitly stated in texts, but results from the interpretation (appraisal) of the actions contained in the situation described, as well as the properties of their actors and objects. Thus, we set up a framework for representing situations described in text as chains of actions (with their corresponding actors and objects), and their corresponding properties (including the affective ones), according to commonsense knowledge. We show the manner in which the so-called “appraisal criteria” can be automatically detected from text and how additional knowledge on the properties of the concepts involved in such situations can be imported from commonsense knowledge bases. Finally, we demonstrate through an extensive evaluation that such a representation is useful to obtain an accurate label of the emotion expressed in text, without any linguistic clue being present therein, increasing the recall of systems performing sentiment analysis from texts.

1.2. MOTIVATION

The radical shift in the method employed for communication and the content of this communication has brought with itself new challenges, but also many opportunities.

At the economic level, the globalization of markets combined with the fact that people can freely express their opinion on any product or company on forums, blogs or e-commerce sites led to a change in the companies’ marketing strategies, in the rise of awareness for client needs and complaints, and a special attention for brand trust and reputation. Specialists in market analysis, but also IT fields such as Natural Language Processing, demonstrated that in the context of the newly created opinion phenomena, decisions for economic action are not only given by factual information, but are highly affected by rumors and negative opinions. Wright

(2009) claims that “for many businesses, online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace”[1]. Studies showed that financial information presented in news articles have a high correlation to social phenomena, on which opinions are expressed in blogs, forums or reviews. On the other hand, many tasks that involved extensive efforts from the companies’ marketing departments are easier to perform. An example is related to market research for advertising, business intelligence and competitive vigilance. New forms of expression on the web made it easier to collect information of interest, which can help to detect changes in the market attitude, discover new technologies, machines, markets where products are needed and detect threats. On the other hand, using the opinion information, companies can spot the market segments their products are best associated with and can enhance their knowledge on the clients they are addressing and on competitors. The analysis of the data flow on the web can lead to the spotting of differences between the companies’ products and the necessities expressed by clients and between the companies’ capacities and those of the competitors. Last, but not least, the interpretation of the large amounts of data and their associated opinions can give companies the capacity to support decision-making through the detection of new ideas and new solutions to their technological or economic problems.

The opinionated data on the web has also produced important changes in the manner in which communities are able to participate in the elaboration of laws and policies. Consultations with communities that in the past were made through the use of questionnaires are now easily made through forums of opinion. Additionally, such data on the opinions that people have about laws, policies, and administrators can be extracted from a variety of sources (e.g. microblogs, blogs, social networks).

The advantage and, at the same time, issue related to these new capabilities is the large amount of information available and its fast growing rate. Lack of information on markets and their corresponding social and economical data, leads to wrong or late decisions and finally to important financial losses. Lack of information of the policy makers leads to wrong decisions, affecting large communities.

Although mostly positive, there are also downsides to the increasing communication through the use of Web 2.0 technologies. The development of social networks and communication between their members led to the development of interesting phenomena, whose effects are both positive and negative and which are difficult to assess. Within social networks gathered around the most peculiar topics, people talk about subjects that they would not address in their everyday life and with their friends or family. Under the hidden identity on the web, however, they are free to express their innermost fears and desires. That is why, allowing and

[1] www.nytimes.com/2009/08/24/technology/internet/24emotion.html?_r=1&ref=start-ups

supporting free communication led to the birth of sites where violence is predicated and encouraged, where people with psychological problems or tendencies towards suicide, addictions etc. talk to one another and encourage their negative behaviors. Such sites must be discovered and controlled, in order to keep under control different social issues that may arise from the described potentially conflictive situations.

As we can see, the present reality is profoundly marked by the opinionated information present both in traditional, as well as new textual genres. Given the proven importance of such data, but also the challenges it raises, in terms of volume, automatic systems, sentiment analysis has become a highly active research field in NLP in the past years.

In the next section, we will show that this task is not only useful in order to obtain important knowledge from non-factual data, but that it also contributes to the improvement of systems dealing with other NLP challenges.

1.3. APPLICATIONS

In the Motivation Section, we explained the main reasons for doing research in the field of sentiment analysis and the applications it has in real-world scenarios. Further on, we will present some of the fields and domains in which this task is useful.

Research has been conducted in the field of opinion mining, aimed at improving different social, economical, political and psychological aspects of every-day human life. There are many applications of opinion mining systems to real-world scenarios. Some of these applications are already available online, others are still under research and other directions and developments in the field are merely appearing. There are sites like “swotty.com”, which mine, classify and summarize opinions from reviews on products on the e-commerce sites that people can use for comparison, advice or recommendation. Other applications of the task, directly related to the commerce and competition markets of companies, use opinion mining from the web to obtain direct, sincere and unbiased market feedback, about their own products (business intelligence), and of the products of their market competition (competitive vigilance). Companies, as well as public figures, use opinion mining to monitor their public image and reputation (trust). Authors can benefit from opinion mining to track their literary reputation.

It was demonstrated that fluctuation in public opinion correlates to fluctuations of stock prices for the targeted companies (Devitt and Ahmad, 2007). Thus, opinion mining can be used to track opinion across time for market and financial studies, for early action in predicted crisis situations or for the issuing of alerts.

Recent developments of the Social Web technologies and the growing number of people writing and reading social media (blogs, forums etc.) also allows for the monitoring and analysis of social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Examples of sites implementing these concepts are “wefeelfine.org” or “twends.com”.

Yet another application of sentiment analysis is the tracking of political view, to detect consistency and inconsistency between statements and action at the government level. It was recently stated that election results could be better predicted by following the discussion threads in blogs.

eRulemaking, as a democratic way of consulting the whole targeted population when a law or policy is to be implemented, can also highly benefit from sentiment analysis, as method to spot and classify a large quantity of subjective data. This task is also performed when tracking views on laws from legal blogs (blawgs).

Last, but not least, studying affect related phenomena is basic for Human-Computer Interaction (Picard, 1995), as most reactions and interactions are not only rationality-based, but heavily rely on emotion.

It was also demonstrated that opinion mining improves other Natural Language Processing tasks, such as:

- Information Extraction, by separating facts from opinions (Riloff et al., 2005);
- Question Answering (Somasundaran et al., 2007), where the application of opinion mining can improve the answering of definition questions (Lita et al., 2005)
- and Multi-Perspective Question Answering (Stoyanov et al., 2005; Yu and Hatzivassiloglou, 2003) where there is not a single, true and correct answer, but a set of answers describing the attitude of different persons on a given fact;
- Summarization of multi-perspective texts (Ku et al., 2005; Ku et al., 2006), where redundant opinion (opinion of the same polarity, given the same arguments) must be removed;
- Authorship (source) determination (Teufel and Moens, 2000; Piao et al., 2007);
- Word Sense Disambiguation (Wiebe and Mihalcea, 2006).

The next chapters of this thesis present different methods and approaches for tackling the task of sentiment analysis in different text types, languages and in the context of a variety of final applications, addressing the five research questions we described. First of all, however, in order to ensure an understanding of the terminology we will employ, in Chapter 2 we present an overview of the tasks and related concepts definition. The main motivations for defining the concepts involved in this task is that the issues related to the study of affective phenomena have been studied for a long time in disciplines such as Psychology or Philosophy and that sentiment analysis in NLP is a recent field, in which the terminology is not yet fully established.

CHAPTER 2. TASKS AND CONCEPTS

Motto: *"How much has to be explored and discarded before reaching the naked flesh of feeling?"(Claude Debussy)*

Having seen the high number of practical applications to the automatic processing of subjective and opinionated language, it is of no wonder that the tasks of subjectivity and sentiment analysis have registered a growing interest from the NLP research community in the past few years. Vast amounts of research has been performed, all falling within the scope of developing computational methods for text analysis, in order to discover whether it is subjective or objective, whether it contains opinions, sentiments, attitudes (and if so, what polarity/tonality/orientation these have) or emotions.

Due to the large number of applications in the social, political and economic spheres, most work has concentrated on creating and evaluating methods, tools and resources to discover whether a specific "object" (person, product, organization, event, etc.) is "regarded"[2] in a positive or negative manner by a specific "source" (i.e. a person, an organization, a community, people in general, etc.). This task has been given many names, from opinion mining, to sentiment analysis, review mining, attitude analysis, appraisal extraction and many others. At the same time, the term "opinion mining", which authors such as Pang and Lee (2008) consider being equivalent to "sentiment analysis", has been employed to denote work that aims at classifying text according to different criteria:

a) the polarity of the sentiment expressed (into positive and negative; sometimes the neutral category is also employed);
b) whether the text includes good or bad news (Ku et al., 2005);
c) whether the candidate that the text is talking about is likely or unlikely to win (Kim and Hovy, 2005);
d) whether the text is expression support or opposition (Bansal et al., 2008; Terveen et al., 1997) ;
e) pros and cons (Kim and Hovy, 2006);
f) determining the polarity of the outcome (e.g. improvement versus death in medical texts) (Niu et al., 2005)
g) Whether a person agrees or disagrees with a topic (political debates) (Balahur et al., 2009e).

[2] "Regard" is a term we deliberately use here, in order to avoid employing any of the terminology used so far, which is defined and detailed in this chapter. In this context, it refers to: 1) an assessment based on a set of criteria (i.e. personal taste, convenience, social approval, moral standards etc.) or 2) on the emotional effect it has on the person as a cause of this assessment.

Given the high dynamics of the field and the vast amount of research done within its framework in the past few years, the terminology employed in defining the different tasks dealing with subjectivity, as well as the concepts involved in them is not yet uniform across the research community.

In order to establish the scope of our research and ensure the understanding of the approaches presented, we first give an overview of the definitions given to the concepts involved outside the area of Natural Language Processing. Subsequently, we present some of the tasks proposed in the research, the concepts involved in them and the definitions they were given. Finally, we propose a set of definitions to the concepts we will use across this thesis and of the different tasks aiming at the automatic processing of subjective texts.

2.1. SUBJECTIVITY

In Philosophy, ***subjectivity*** refers to the subject and his or her perspective, feelings, beliefs, and desires. (Solomon, 2005).

In NLP, the most widely used definition is the one proposed by Wiebe (1994). The author defines ***subjectivity*** as the “linguistic expression of somebody’s opinions, sentiments, emotions, evaluations, beliefs and speculations”. In her definition, the author was inspired by the work of the linguist Ann Banfield (Banfield, 1982), who defines as subjective the “***sentences that take a character’s point of view*** (Uspensky, 1973)” and that ***present private states*** (Quirk, 1985) (that are not open to objective observation or verification) of an experiencer, holding an ***attitude***, optionally towards an object. Subjectivity is opposed to objectivity, which is the expression of facts. Wiebe et al. (2005) considers the term **private state**, which is in a pragmatic sense equivalent to subjectivity and that is defined as a general term that covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments. According to the definition proposed by Wiebe (1994), an example of subjective sentence is *“This book is amazing!”,* whereas an example of objective sentence is *“This book costs 10€ on Amazon.com.”*

In view of the given definition, the Multi-Perspective Question Answering corpus (Wiebe et al., 2005) takes into account three different types of elements for the annotation of subjectivity: ***explicit mentions of private states*** (e.g. *“The U.S.* ***fears*** *a spill-over,” said Xirao-Nima*), ***speech events expressing private states*** (e.g. *“The U.S. fears a spill-over,”*__said__ *Xirao-Nima*), ***expressive subjective element*** (e.g. *“The report is* **full of absurdities**,*”*).

In the Handbook of Natural Language Processing (2010), Bin Liu defines subjective versus objective sentences as follows: *“An objective sentence expresses* *some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs."*

2.2. OPINION, SENTIMENT, EMOTION. COMPUTATIONALLY-RELATED TASKS.

In the case of opinion, if one were to look at the term definition given in the Webster dictionary[3], they would find the following set of synonyms: "opinion", "view", "belief", "conviction", "persuasion", "sentiment", meaning "a judgment one holds as true". Out of this definition, it is important to stress upon the fact that these closely related, synonym terms, have slightly different meanings.

- ***Opinion*** implies a conclusion thought out yet open to dispute; it is:
    1. A): a view, judgment, or appraisal formed in the mind about a particular matter; B): approval, esteem;
    2. A): a belief stronger than impression and less strong than positive knowledge; B): a generally held view;
    3. A): a formal expression of judgment or advice by an expert; B): the formal expression (as by a judge, court, or referee) of the legal reasons and principles upon which a legal decision is based.
- ***View*** suggests a subjective opinion.
- ***Belief*** implies often deliberate acceptance and intellectual assent.
- ***Conviction*** applies to a firmly and seriously held belief.
- ***Persuasion*** suggests a belief grounded on assurance (as by evidence) of its truth.
- ***Sentiment*** suggests a settled opinion reflective of one's **feelings.**

The term **feeling** is defined as the conscious subjective experience of emotion. (Van den Bos, 2006). This is approximately the same definition as the one given by Scherer (2005), which states that *"the term feeling points to a single component of emotion, denoting the subjective experience process, and is therefore only a small part of an emotion".*

This definition suggests that there are different types of opinions and that not all opinions are subjective (see the definition of "view"), as well as not all opinions have a sentiment associated to them. An "objective" opinion could be considered to be the one of an expert (e.g. a doctor giving a diagnosis on the basis of observed symptoms). A "subjective" opinion is one that is based on personal criteria (depends on the individual taste, ideas, standards etc.). This same definition also pinpoints to the fact that sentiments are types of opinions, namely the ones that are "reflective of one's feelings", where "feeling" is the "conscious subjective

[3] http://www.merriam-webster.com/

experience of emotion". Thus, sentiment relates to emotion, in the sense that it is the expression of an evaluation based on the emotion the writer feels.

"Opinion mining", as a computational task, appeared for the first time in a paper by Dave et al. (2003), and it was defined as follows: *"Given a set of evaluative text documents D that contain opinions (or sentiments) about an "object" (person, organization, product etc.), opinion mining aims to extract attributes and components of the object that have been commented on in each document d in the set D and to determine whether the comments are positive, negative or neutral."* According to Pang and Lee (2008), the fact that this work appeared in the proceedings of the World Wide Web (WWW) 2003 conference explains the popularity of this terminology within the web search and retrieval research community. This also explains the fact that Esuli and Sebastiani (2006) define ***opinion mining*** *as "a recent discipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses".*

From the computational point of view, Kim and Hovy (2005) define **opinion** *"as a quadruple [Topic, Holder, Claim, Sentiment] in which the Holder believes a Claim about the Topic, and in many cases associates a Sentiment, such as good or bad, with the belief.* As far as sentiments are concerned, the authors define them as: *"Sentiments, which in this work we define as an explicit or implicit expression in text of the Holder's positive, negative, or neutral regard toward the Claim about the Topic. Sentiments always involve the Holder's emotions or desires, and may be present explicitly or only implicitly."*

This definition relates opinion with sentiment, in the sense that it states that some opinions carry a sentiment, while others do not. In order to illustrate the difference between opinions with sentiment and opinions without sentiment, Kim and Hovy (2005) provide the following examples:

(1) "I believe the world is flat."

(2) "The Gap is likely to go bankrupt."

These are sentences that express opinions, but they do not contain any sentiment. The following examples, taken from the same paper, explain the difference between explicitly versus implicitly expressed sentiment of opinions:

(3) "I think that attacking Iraq would put the US in a difficult position." (implicit)

(4) "The US attack on Iraq is wrong." (explicit)

(5) "I like Ike." (explicit)

(6) "We should decrease our dependence on oil." (implicit)

Another definition of the term **opinion** was given by Bing Liu (2010). The author is the one who defined the task of "feature-based opinion mining and summarization”, which deals with the classification of opinions expressed on different features of products and their summarization (Hu and Liu, 2004).

According to Liu (2010):

- *“An* **opinion** *on a feature f is a positive or negative view, attitude, emotion or appraisal on f from an opinion holder.”*
- *“The* **holder of an opinion** *is the person or organization that expresses the opinion.”*
- *“An* **explicit opinion** *on feature f is an opinion explicitly expressed on f in a subjective sentence.”*
- *“An* **implicit opinion** *on feature f is an opinion on f implied in an objective sentence.”*
- *“An* **opinionated sentence** *is a sentence that expresses explicit or implicit positive or negative opinions. It can be a subjective or objective sentence.”*
- *“***Emotions** *are our subjective feelings and thoughts.”*

All tasks defined within opinion mining aim at classifying the texts according to the “orientation of the opinion” (usually into three classes – of positive, negative and neutral). The classes of opinion considered have been denoted using different terms: ***opinion orientation, sentiment polarity, polarity, sentiment orientation, polarity of opinion, semantic orientation*****.**

As far as sentiment analysis as NLP task is concerned, most of the research in the field coincides with the following definition: *“The binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion is called sentiment polarity classification or polarity classification”.* (Pang and Lee, 2008)

“The orientation of an opinion on a feature f indicates whether the opinion is positive, negative or neutral. Opinion orientation is also known as sentiment orientation, polarity of opinion, or semantic orientation.”(Liu, 2010)

A related concept is ***valence***, defined as *“a negative or positively attitude”* (Polanyi and Zaenen, 2004). In relation to this concept, Polanyi and Zaenen (2004) define the so-called ***“contextual valence shifters”*** *(e.g. negatives and intensifiers, modals, presuppositional items, ironical formulations, connectors),* which are lexical items or formulations that change the orientation of the attitude.

The term ***“sentiment”*** in the context of a computational text analysis task is mentioned for the first time in the paper by Das and Chen (2001). According to the authors *“in this paper, ‘sentiment’ takes on a specific meaning, that is, the net of positive and negative opinion expressed about a stock on its message board.”* . At the same time, Tong (2001) proposed a “new” task at the Workshop on Operational Text Classification (OTC2001), which concerned the detection and tracking of opinions in on-line discussions and the subsequent classification of the sentiment of opinion.

The aim of the paper by Turney (2002) is *"to classify reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. A phrase has a positive semantic orientation when it has good associations (e.g., "subtle nuances") and a negative semantic orientation when it has bad associations (e.g., "very cavalier")".*

Pang et al. (2002) propose different methods to determine the *"sentiment, or overall opinion towards the subject matter for example, whether a product review is positive or negative".*

Nasukawa and Yi (2003) entitled their paper, "Sentiment analysis: Capturing favorability using natural language processing". In this paper, they state that *"the essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject."*

Yi et al. (2003), in their paper "Sentiment Analyzer: Extracting sentiments about a given topic using natural language processing techniques", consider opinion an equivalent term to sentiment. Their approach approximates the task later known as "feature-based opinion mining and summarization" (Hu and Liu, 2004), as they extract sentiment in correlation to a specific topic.

Subjectivity analysis and sentiment analysis/opinion mining have been considered to be highly-related tasks. Pang and Lee (2003) state that subjectivity analysis performed prior to sentiment analysis leads to better results in the latter. Banea et al. (2010) states in this sense that *"while subjectivity classification labels text as either subjective or objective, sentiment or polarity classification adds an additional level of granularity, by further classifying subjective text as either positive, negative or neutral".*

However, according to Pang and Lee (2008): *"(...) nowadays many construe the term (sentiment analysis) more broadly to mean the computational treatment of opinion, sentiment, and subjectivity in text."*

As we can observe, terminology employed in this field is highly variable. At times, the definitions used to denote one task or another and their related concepts are vague, inexact, overlap with definitions given for other terms, different terms are used to denote the same task or concept and the definitions are not consistent with the formal ones (that we can find, for example, in a dictionary). On top of their inconsistencies, there is also a large body of research performing emotion detection to improve sentiment analysis (Cambria et al., 2009), although no explicit relation between emotion, sentiment and opinion is presented.

We will further on define emotion and relate it to the wider context of affect and the term “affective computing”, used in Artificial Intelligence. Subsequently, we clarify the connection between all these concepts, within the frame of the Appraisal Theory[4].

Affect is “a superordinate concept that subsums particular valenced conditions such as emotions, moods, feelings and preferences” (Ortony et al., 2005), being one of the four components whose interaction make the human organism “function effectively in the world” (Ortony et al., 2005), along with motivation, cognition and behaviour.

Affective computing is a branch of Artificial Intelligence (AI) dealing with the design of systems and devices that can recognize, interpret, and process human affect. The concept includes interdisciplinary work from computer science, but also psychology and cognitive science. However, the term was introduced in the study within AI by (Picard, 1995), who envisages both the capabilities of computers to interpret affect from digital content, as well as imitate affect in humans.

Emotion is a complex phenomenon, on which no definition that is generally accepted has been given. However, a commonly used definition considers emotion as “an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems (Information processing, Support, Executive, Action, Monitor) in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism”. (Scherer, 1987; Scherer, 2001).

Emotion detection and classification is the task of spotting linguistic expressions of emotion from text and classifying them in predefined categories/labels (e.g. anger, fear, sadness, happiness, surprise, disgust etc.). Emotion is a much more complex phenomenon, whose expression in language may not always have a subjective form (Ortony, 1997).

Let us consider a few examples, to note the difference between subjectivity, opinion, emotion and sentiment, in a sense that is consistent to the definitions these concepts are given outside the NLP world. In the following table, “Y” corresponds to “yes”, “N” corresponds to “no”, “C” corresponds to “context” (dependence on the context), “POS” to “positive”, “NEG” stands for “negative” and “NEU” for “neutral”. The symbol “---“ stands for the lack of the corresponding element.

[4] This set of theories have been proposed in Psychology, by De Rivera (1977), Frijda (1986), Ortony, Clore and Collins (1988), Johnson-Laird and Oatley (1989). It has also been used in to define the Appraisal Framework in Linguistics, by Martin and White(2001).

Nr.	Example	Subjective?	Opinionated?	Emotion?	Sentiment?	Polarity? POS/NEG/NEU (if there is a sentiment)
1.	It broke in two days.	N	N	Y	Y	NEG
2.	It's great!	Y	Y	Y	Y	POS
3.	The bank is likely to go bankrupt.	Y	Y	C	C	---
4.	The president denied the existence of any financial crisis during the election campaign.	N	C	C	C	NEG/--- (C)
5.	The screen is really huge!	Y	Y	Y	Y	POS/NEG/(C)
6.	It took them three years to fix the puthole in front of our building.	N	Y	Y	Y	NEG
7.	They always advertise this product as cheap and chic. Can you actually say that?	Y	Y	Y	Y	NEG
8.	These headphones are perfect. If you are deaf, that is.	Y	Y	Y	Y	NEG
9.	I am firmly convinced they came yesterday.	Y	Y	C	C	POS/NEG/NEU (C)
10.	This car costs 14.000 Euros.	N	C	C	C	POS/NEG/NEU (C)
11.	This car only costs 14.000 Euros.	Y	Y	C	Y	POS
12.	The president should prepare his arguments for the next negotiations.	Y	Y	C	C	NEG/NEU (C)
13.	It is rumored he might sell his business.	N	Y	N	N	---
14.	This book costs 10€ on Amazon.com	N	N	N	N	---
15.	On Amazon, this book costs 10€!	N	Y	Y	Y	POS/NEG (C)
16.	He killed all the flies.	N	N	N	N	---
17.	It killed all the birds.	N	Y	Y	Y	NEG
18.	In the end, he killed all the bad guys and walked into the sunset with the girl.	N	C	C	C	POS/NEG (C)

Table 2.1: Examples illustrating the concepts of sentiment, emotion, opinion, subjectivity and objectivity

As we can see, these classes that are very frequently used in opinion mining/sentiment analysis and subjectivity analysis are not related in a straightforward manner. As stated by some of the definitions we presented, not all opinions are neither necessarily subjective in nature, nor do they have to contain sentiment in all the cases. Sentiments on different targets can also be conveyed by presenting arguments that are factual in nature (e.g. “The new phone broke in two days”), as the underlying emotion that is expressed is not always stated directly. Moreover, subjective statements that contain an opinion must not necessarily contain an emotion, thus they have no sentiment correlated to them (e.g. “I believe in God”). Finally, a text may express and emotion without expressing any opinion and sentiment (e.g. “I’m finally relaxed!”). This idea can be summarized in the following schema:

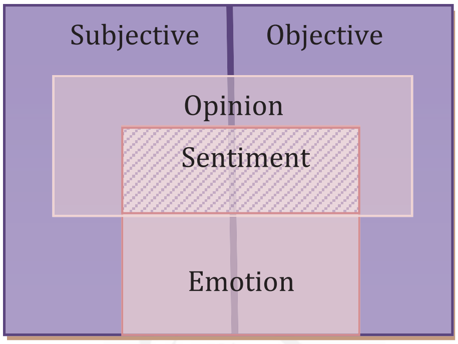


Figure 2.1: The relation between subjectivity, objectivity, opinion, sentiment and emotion

2.3. ATTITUDE AND APPRAISAL. APPRAISAL THEORIES.

Although the definitions of the tasks, as well as of the concepts they involve are highly discrepant, all the tasks which are concerned with opinion, sentiment, attitude, appraisal or emotion and which were denominated “sentiment analysis” or “opinion mining” (and the aforementioned related terms) actually have the same aim. This can be summarized, in a very naïve manner, by the statement *“Does the source of the text like/appreciate/think positively/feel good/is happy/ about the target of what is stated or not?”; “What exactly does the source like/appreciate/think positively/feel good/is happy or does not like/appreciate/think positively/feel good/is happy about?”; “How does he/she express that?”*.

It is very interesting to note that most of the literature does not consider any further elements of the act of communication – the source of what is said and who it is intended for. The Speech-Act theory (Austin, 1976) states that “When we speak, our words do not have meaning in and of themselves. They are very much affected by the situation, the speaker and the listener. Thus words alone do not have a simple fixed meaning.” Therefore, apart from what is explicitly stated in a text, it is very important to take into consideration the ***context of what is said*** (who is saying it, why, what is the ***intention behind what he/she is saying***; is the act of writing purely to inform the reader on a specific "object", is it to produce a specific emotion to the reader, is it to convince him/her about something, is it to purely express his/her own regard on this "object" he/she is writing about?) and ***whom it is addressed to*** (would a potential reader like what he/she is reading, would he be comfortable with it, would he resent it, would he think it is good, positive, would he feel happy about it).

Following on this idea and based on the definitions we have seen so far, we can claim that much of the work that has been done under the umbrella of "sentiment analysis" or "opinion mining" is actually concerned with detecting "attitudes" and classifying them, since, in fact, the act of communication (in this case, through text) is intended for a reader. Both the writer (with his own views on the world), as well as the reader, who is not a mere passive entity, but who actively uses his/her own knowledge of the world, the context and his/her affect to interpret what is written should be taken into consideration at the time of deciphering the sentiments expressed in text. Additionally, especially in the traditional textual genres such as newspaper articles, where writers are thought to be objective when rendering a piece of news, expressions of sentiment cannot be direct. Opinions are expressed in a non-subjective manner, by omitting certain facts and overly repeating others (Balahur and Steinberger, 2009).

The work in the field of NLP that deals with the computational treatment of attitude is called ***"attitude analysis"*** or ***"appraisal analysis"*** relating it to the Appraisal Theory**.**

An **attitude** (Breckler and Wiggins, 1992) is a "hypothetical construct that represents an individual's degree of like or dislike for something. Attitudes are generally positive or negative views of a person, place, thing, or event— this is often referred to as the attitude object. People can also be conflicted or ambivalent toward an object, meaning that they simultaneously possess both positive and negative attitudes toward the item in question. Attitudes are judgments. They develop on the **ABC** model (affect, behavior, and cognition). The *affective* response is an emotional response that expresses an individual's degree of preference for an entity. The *behavioral* intention is a verbal indication or typical behavioral tendency of an individual. The *cognitive* response is a cognitive evaluation of the entity that constitutes an individual's beliefs about the object. Most attitudes are the result of either direct experience or observational learning from the environment."

Work that has concentrated on attitude analysis was done by Taboada and Grieve (2004), Edmonds and Hirst (2002), Hatzivassiloglou and McKeown (1997) and Wiebe et al. (2001). However, only the first work considered "attitude" as different from “subjectivity”, although subjectivity, as we have seen, does not always imply an evaluation as in the case of attitude.

The work by Taboada and Grieve (2004) and the recent work by Neviarouskaya et al. (2010) also try to link the concept of sentiment (judgment/appreciation) with attitude, based on the Appraisal Theory (Martin and White, 2005). This is a framework of linguistic resources inscribed in discourse semantics, which describes how writers and speakers express inter-subjective and ideological positions (the language of emotion, ethics and aesthetics), elaborating on the notion of interpersonal meaning, i.e. social relationships are negotiated through evaluations of the self, the others and artifacts. According to this theory, emotion is achieved by appraisal, which is composed of ***attitude*** (affect, appreciation, and judgment), ***graduation*** (force and focus), ***orientation*** (positive versus negative) and ***polarity*** (which can be marked or unmarked).

Only the work by Neviarouskaya et al. (2010) distinguishes among the different components of attitudes and employs different methods to tackle each of the issues in this context (i.e. direct and indirect expressions of sentiments, through judgments and appreciations or emotions, using emotion detection and sentiment analysis techniques). Nonetheless, this work, too, remains at a lexical level of analysis.

Although the terminology employed in defining the tasks related to the computational treatment of subjectivity and sentiment is still not well-established, the body of research performed in the field makes it impossible to draw a line between what could be wrongfully defined and what is correctly defined.

The tasks of subjectivity analysis, opinion mining or sentiment analysis, as we will see along the following chapters, are defined and tackled very differently depending on the final aim of the application, the type of text on which it is applied and the context in which it is performed. As a consequence, in the next chapters, besides the description of the tasks we aim at resolving using sentiment analysis, we will also describe each of the definitions given in the context of the tasks sentiment analysis was applied to.

The operational definition we will use for the concept of “sentiment” along this thesis will combine the understanding that is given by the literature to sentiment and attitude together. Thus, we consider ***sentiment*** as a settled opinion reflective of one’s ***feelings*** *“a single component of emotion, denoting the* ***subjective*** *experience process” (Scherer, 2005),* implicitly or explicitly present in text through expressions of affect, appreciation, judgment, but also an *expression of behavior and cognition*. In this context, sentiments are not only present in subjective sentences, but can also be expressed in objective sentences (e.g. “It broke in two days”, implicitly describing a negative appreciation of the quality of the product described). In the context where the term opinion is employed, we refer to the specific type of opinions represented by sentiments (defined as above).

CHAPTER 3. STATE OF THE ART

Motto: *”Compelling reason will never convince blinding emotion.”(Richard Bach)*

Research in affect has a long established tradition in many sciences - linguistics, psychology, socio-psychology, cognitive science, pragmatics, marketing or communication science. Recently, as we have seen in the previous chapter, many closely related subtasks were developed also in the field of affect computing, from emotion detection, subjectivity analysis, opinion mining to sentiment analysis. As we will show in the next sections and chapters, this research area covers many aspects in Artificial Intelligence (AI) and Natural Language Processing (NLP), each with its challenges and proposed solutions.

Not less important is the fact that subjectivity analysis, for example, was demonstrated to bring substantial improvement over other NLP tasks, such as information extraction (IE), question answering (QA), authorship determination, text categorization, multi-document summarization and was employed as a filtering step for sentiment analysis. Opinion mining (sentiment analysis), on the other hand, besides being employed per se in order to seek attitude towards different entities, has become a crucial component in Natural Language Processing systems that answer mixed types of questions or summarize texts pertaining to new, emerging text types (blogs, e-commerce sites reviews, forums etc.).

In this chapter, we present a general overview of the state of the art methods and resources that were employed and created for the tasks of subjectivity and sentiment analysis. Additionally, we present the state of the art methods employed in general affect detection challenges and their application in the context or jointly with other NLP tasks. We start by presenting the resources created for the task of subjectivity analysis.

3.1. SUBJECTIVITY ANALYSIS

The study of subjectivity in text from a computational linguistics point of view began together with the work of Janyce Wiebe (Wiebe, 1994), who, based on Ann Banfield’s (Banfield, 1982) linguistic theories on narrational aspects and Quirk’s definition, centered the idea of subjectivity around that of “private states” (Quirk, 1985), events which are not open to objective observation and verification, the linguistic expression of somebody’s opinions, sentiments, emotions, evaluations, beliefs, speculations (Wiebe et al., 2004). The aim in her research was to recognize opinion-oriented language, which she denominated as “subjectivity indicators”, and to distinguish it from objective descriptions.

The aim of subjectivity analysis, as described in the task definition, is to classify content into objective or subjective. This task is not trivial, as many expressions carry in themselves a certain subjectivity (which we will later show is related to the content and text intentionality and many expressions are used both in a subjective, as well as objective manner. An example of factual text would be “100 people died after a bomb exploded in the centre of the town” and an example of subjective text is “Hundreds of prisoners were mercilessly tortured in the prison”.

Research in the field of subjectivity analysis concentrated on the creation of resources and the definition of methods to detect subjectivity indicators.

3.1.1. RESOURCE CREATION FOR SUBJECTIVITY ANALYSIS

A) MANUALLY CREATED RESOURCES

Resources containing subjective words are “The General Inquirer”[5] (Stone et al., 1966), which was not specifically designed for subjectivity analysis, but contains 11788 sense disambiguated words, out of which the subjective ones are annotated correspondingly with polarity, strength and according to axes of emotion. The resource was created manually.

Comlex (Macleod et al., 1994) is a dictionary containing 38000 words for the English language that was specifically built for NLP systems. This dictionary also contains a large number of attitude adverbs.

Another resource for subjectivity analysis, which is also annotated as far as polarity is concerned, is the Multi-Perspective Question Answering (MPQA) corpus[6] (Wiebe et al., 2005; Wiebe and Wilson, 2005; Wilson and Wiebe, 2003), containing annotations of 10000 sentences from the world press. The annotation scheme is complex, with the intention of spotting as many of the subjectivity and emotion-related aspects as possible. Subjectivity is defined around “private states” (Wiebe, 1994). For the MPQA annotation process, each of the subjectivity expressions is defined within a private state frame (Wiebe et al., 2005), which includes the source of the private state (which can be “nested”, i.e., it is not only the author that can be the source of the private state described, but the author may introduce the private states of another person), the target and properties involving the intensity of the subjectivity expressed, the significance and the type of attitude. The second annotated type of element is “objective speech event”, which, in

[5] http://www.wjh.harvard.edu/~inquirer/

[6] http://www.cs.pitt.edu/mpqa/

contrast to the private frames, contain annotations of events that are objective in nature.

The Opinion Finder lexicon (subjectivity clues) (Wilson et al., 2005) contains over 8000 words, annotated also at the level of polarity value, and was built starting with the grouping of the subjectivity clues in (Riloff and Wiebe, 2003) and enriched with polarity annotated subjective words taken from the General Inquirer and the lexicon proposed by Hatzivassiloglou and McKeown (1997). It is interesting to notice that authors found most of the subjective words to have either a positive or a negative polarity and only few were both positive and negative or neutral.

B) SEMI-AUTOMATICALLY AND AUTOMATICALLY CREATED RESOURCES

Another annotation scheme and corpus for subjectivity versus objectivity classification, as well as polarity determination at sentence level was developed by Yu and Hatzivassiloglou (2003), in a semi-automatic manner. The authors start from a set of 1336 seed words, manually annotated by Hatzivassiloglou and McKeown (1997), extended by measuring co-ocurrence between the known seed words and new words. The hypothesis on which the authors based their approach is that positive and, respectively, negative words, tend to co-occur more than it is expected by chance. As measure for association, the authors employ log-likelihood on a corpus that is tagged at the part-of-speech level.

A resource for subjectivity was built semi-automatically on the basis of the Appraisal Theory (Martin and White, 2005). The Appraisal Theory is a framework of linguistic resources inscribed in discourse semantics, which describes how writers and speakers express inter-subjective and ideological positions (the language of emotion, ethics and aesthetics), elaborating on the notion of interpersonal meaning, i.e. social relationships are negotiated through evaluations of the self, the others and artifacts. According to this theory, emotion has achieved by appraisal, which is composed of attitude (affect, appreciation, judgment), graduation (force and focus), orientation (positive versus negative) and polarity (which can be marked or unmarked). A lexicon of appraisal terms is built by Whitelaw et al. (2005), based on the examples provided by Martin and White (2005) and Matthiassen (1995) (400 seed terms) and patterns in which filler candidates were extracted from WordNet (Fellbaum ed., 1999). Term filtering was done by ranking obtained expressions and manually inspecting terms that were ranked with high confidence. The resulting lexicon contains 1329 terms.

C) APPROACHES TO MAPPING SUBJECTIVITY RESOURCES TO OTHER LANGUAGES

Most of the work in obtaining subjectivity lexicons was done for English. However, there were some authors who developed methods for the mapping of subjectivity lexicons to other languages.

To this aim, Kim and Hovy (2006) use a machine translation system and subsequently use a subjectivity analysis system that was developed for English to create subjectivity analysis resources in other languages.

Mihalcea et al. (2007) propose a method to learn multilingual subjective language via cross-language projections. They use the Opinion Finder lexicon (Wilson et al., 2005) and use two bilingual English-Romanian dictionaries to translate the words in the lexicon. Since word ambiguity can appear (Opinion Finder does not mark word senses), they filter as correct translations only the most frequent words. The problem of translating multi-word expressions is solved by translating word-by-word and filtering those translations that occur at least three times on the Web.

Another approach in obtaining subjectivity lexicons for other languages than English was explored by Banea et al. (Banea et al., 2008b). To this aim, the authors perform three different experiments, obtaining promising results. In the first one, they automatically translate the annotations of the MPQA corpus and thus obtain subjectivity annotated sentences in Romanian. In the second approach, they use the automatically translated entries in the Opinion Finder lexicon to annotate a set of sentences in Romanian. In the last experiment, they reverse the direction of translation and verify the assumption that subjective language can be translated and thus new subjectivity lexicons can be obtained for languages with no such resources.

Further on, another approach to building lexicons for languages with scarce resources is presented by Banea et al. (Banea et al., 2008a). In this research, the authors apply bootstrapping to build a subjectivity lexicon for Romanian, starting with a set of seed subjective entries, using electronic bilingual dictionaries and a training set of words. They start with a set of 60 words pertaining to the categories of noun, verb, adjective and adverb from the translations of words in the Opinion Finder lexicon. Translations are filtered using a measure of similarity to the original words, based on Latent Semantic Analysis (LSA) (Deerwester et al., 1990) scores.

Yet another approach to mapping subjectivity lexica to other languages is proposed by Wan (2009), who uses co-training to classify un-annotated Chinese reviews using a corpus of annotated English reviews. He first translates the English reviews into Chinese and subsequently back to English. He then performs co-training using all generated corpora.

Kim et al. (2010) create a number of systems consisting of different subsystems, each classifying the subjectivity of texts in a different language. They translate a corpus annotated for subjectivity analysis (MPQA), the subjectivity clues (Opinion finder) lexicon and re-train a Naïve Bayes classifier that is implemented in the Opinion Finder system using the newly generated resources for all the languages considered.

Finally, Banea et al. (2010) translate the MPQA corpus into five other languages (some with a similar ethimology, others with a very different structure). Subsequently, they expand the feature space used in a Naïve Bayes classifier using the same data translated to 2 or 3 other languages. Their conclusion is that by expanding the feature space with data from other languages performs almost as well as training a classifier for just one language on a large set of training data.

3.1.2. APPLICATIONS OF SUBJECTIVITY ANALYSIS

Automatic subjectivity analysis was proven to be helpful as filtering step in Information Extraction (IE) (Somasundaran et al., 2007), as subjectivity often causes false hits for IE.

Riloff et al. (2005) present a strategy to improve IE through the use of subjectivity. In their approach, the authors use the Opinion Finder subjectivity lexicon in order to create subjectivity patterns, which help them build a classifier that is able to capture feature above the word level and takes into consideration the surrounding context of the subjective expressions. At the time of performing IE, they discard the subjective sentences within the results and obtain a large improvement over previous approaches that do not distinguish sentences from the subjectivity/objectivity point of view.

Pang and Lee (2004) demonstrate that using subjectivity analysis as the first step towards opinion mining yeilds better results.

Stoyanov et al. (2004) show that subjectivity analysis used in the context of separating opinionated answers to multi-perspective questions improves the results in this task.

Wiebe and Mihalcea (2006) research on the effect of subjectivity detection for Word Sense Disambiguation (WSD). They show that distinguishing subjective contexts can help to assign the right sense to words that can be used both as objective as well as subjective. This distinction is also used by Rumbell et al. (2008), where the authors identify the figurative use of language (metaphorical senses of words) to classify sentiment in conversations.

Sentiment analysis, also known as opinion mining (Pang and Lee, 2008) attempts to identify the opinion/sentiment that an entity may hold towards an object and it involves a more profound, fine-grained analysis of the text compared to subjectivity analysis.

Some authors consider that sentiment analysis is subsequent to the task of subjectivity detection (Pang and Lee, 2002). In this sense, sentiment analysis continues with the classification of identified sentiment-containing text into two or three categories (positive, negative or positive, negative and neutral). The neutral category corresponds, in this case, to the objective category in subjectivity analysis, although one may imagine that neutrality can also be expressed as opinions that do not have a clear tendency towards positive or negative.

Although intuitively subjective texts express polar sentiments, Wilson et al. (2005) show that factual pieces of text contain indirectly expressed opinions and opinionated texts also contain factual expressions, as a means of argumentation for an idea. Therefore, sentiment analysis has been approached by the majority of the research independently from subjectivity analysis.

Opinions have three basic components. The first one is the “opinion holder”, which is the “source” of the opinion; the second element is the “object”, which is the “target” of the opinion; finally, the third element is the “opinion” (a view, attitude, or appraisal on an object from an opinion holder), which can be seen as a “private state” concerning the target. There are many subtasks that were defined within the main sentiment analysis task. For each of these, resources were created and different methods were developed. There are three main research areas around which these tools evolved:

- **Creation of resources** for sentiment analysis/ opinion mining;
- **Classification of text** (whose main aim is finding expressions of emotion and classifying the texts into positive and negative), a task which has been performed at a document, sentence, phrase and word level;
- **Opinion extraction** (which is concerned with finding parts of text with opinion, identifying the polarity of the sentiment expressed and determining the source and target of the sentiment expressed).

Recent work has also considered opinion mining as a two-stage problem (Jijkoun et al., 2010), in an attempt to join the two communities that have been working in this field (Information Retrieval and Information Extraction) and offer an end-to-end solution to the opinion analysis problem – from the retrieval to the classification stages. Thus the authors believe that two different, but complementary problems can be identified (Jijkoun et al., 2010):

- ***Sentiment extraction:*** *“given a set of textual documents, identify phrases, clauses, sentences or entire documents that express attitudes, and determine the polarity of these attitudes (Kim and Hovy, 2004)”; and*
- ***Sentiment retrieval:*** *“given a topic (and possibly, a list of documents relevant to the topic), identify documents that express attitudes toward this topic (Ounis et al., 2007)”.*

3.2.1. CREATION OF RESOURCES FOR SENTIMENT ANALYSIS

As we mentioned in Section 3.1, some authors consider that the task of sentiment analysis is subsequent to that of subjectivity analysis, involving an extra step: the classification of the retrieved opinion words according to their polarity. Thus, the existing lexical resources for the opinion task contain words and expressions that are subjective and that have a value for polarity assigned.

At the same time, distinguishing a text that presents facts from one that depicts subjective statements is more straightforward, whereas classifying what is said into valence categories involves a more profound analysis of the context, the structure of the text, the presence of modifier expressions etc.

Thus, apart from creating lexical resources that contain words and expressions with their corresponding a priori assigned polarity, research in sentiment analysis also concentrated on the development of annotation schemes that can be used to label corpora in order to capture the specificities of the different expressions of opinion (be it direct, indirect or implicit), in the diverse types of text from which opinion is mined (news/blogs/product reviews).

It was demonstrated that corpora labeling is a necessary step for the training and evaluation of systems implementing sentiment analysis and that fine-grained opinion mining requires the use of such resources. However, some approaches to document-level sentiment analysis (such as opinion mining of movie reviews) use as gold standard texts that are already “classified”, since the e-commerce sites where they are taken from allow for product reviewers to assign “stars” to the different categories describing a product (from 1 to 5 stars, 1 star being “bad” and 5 stars meaning “very well”).

There are a series of techniques that were used to obtain lexicons of subjective words with associated polarity.

Hu and Liu (2004) start with a set of seed adjectives (“good” and “bad”) and apply synonymy and antonymy relations in WordNet.

A similar approach was used in building WordNet Affect (Strapparava and Valitutti, 2004), starting from a larger set of seed affective words, classified according to the six basic categories of emotion (joy, sadness, fear, surprise, anger and disgust) and expanding the lexicon using paths in WordNet.

Another related method was used in the creation of SentiWordNet (Esuli and Sebastiani, 2005). The idea behind this resource was that "terms with similar glosses in WordNet tend to have similar polarity". Thus, SentiWordNet was build using a set of seed words whose polarity was known and expanded using gloss similarity.

As mentioned in the subjectivity classification section, in the collection of appraisal terms in Whitelaw et al. (2005), the terms also have polarity assigned.

MicroWNOp (Cerini et al., 2007), another lexicon containing opinion words with their associated polarity, was built on the basis of a set of terms (100 terms for each of the positive, negative and objective categories) extracted from the General Inquirer lexicon and subsequently adding all the synsets in WordNet where these words appear. The criticism brought to such resources is that they do not take into consideration the context in which the words or expressions appear. Other methods tried to overcome this critique and built sentiment lexicons using the local context of words.

Pang et al. (2002) built a lexicon of sentiment words with associated polarity value, starting with a set of classified seed adjectives and using conjunctions ("and") disjunctions ("or", "but") to deduce orientation of new words in a corpus.

Turney (2002) classifies words according to their polarity on the basis of the idea that terms with similar orientation tend to co-occur in documents. Thus, the author computes the Pointwise Mutual Information score between seed words and new words on the basis of the number of AltaVista hits returned when querying the seed word and the word to be classified with the "NEAR" operator.

In our work in, we compute the polarity of new words using "polarity anchors" (words whose polarity is known beforehand) and Normalized Google Distance (Cilibrasi and Vitanyi, 2006) scores using as training examples opinion words extracted from "pros and cons reviews" from the same domain, using the clue that opinion words appearing in the "pros" section are positive and those appearing in the "cons" section are negative (Balahur and Montoyo, 2008b; Balahur and Montoyo, 2008d; Balahur and Montoyo, 2008f). Another approach that uses the polarity of the local context for computing word polarity is the one presented by Popescu and Etzioni (2005), who use a weighting function of the words around the context to be classified.

The lexical resources that were created and are freely available for use in the task of opinion mining are:

- WordNet Affect (Strapparava and Valitutti, 2004);
- SentiWordNet (Esuli and Sebastiani, 2006);
- Emotion triggers (Balahur and Montoyo, 2008a);
- MicroWNOp (Cerini et al., 2007);
- General Inquirer (Stone et al., 1966);

- Appraisal terms (annotated according to the principles of the appraisal theory framework in linguistics) (Whitelaw et al., 2005).

On the other hand, there are several specific annotation schemes and corresponding corpora that were created for the affect-related applications in NLP:

- The ISEAR (International Survey on Emotion Antecedents and Reactions) corpus (Scherer and Wallbott, 1997);
- The MPQA (Multi-perspective Question Answering System) corpus;
- The EmotiBlog corpus (Boldrini et al., 2009);
- The TAC (Text Analysis Conference) Opinion Pilot data (TAC 2008) and TREC (Text Retrieval Conference) data (from 2006 to 2010), which consists of annotated texts for the different opinion retrieval specific tasks, on the Blog06 collection;
- The NTCIR MOAT (Multilingual Opinion Analysis Track) data (2007-2010), which contains both monolingual annotated data for opinion mining in English, Chinese and Japanese, as well as cross-lingual analysis data (in MOAT 2010).

The recent tasks proposed at NTCIR MOAT concerning the detection of the opinion holder and opinion target, in addition to the opinion classification, has encouraged the development of new corpora that annotates the opinion source and target, at different granularities. In this sense, apart from the work by Boldrini et al. (2009), the work by Toprak et al. (2010) proposes a model for sentence and expression level annotation of opinions in user-generated discourse.

3.2.2. RESEARCH IN TEXT SENTIMENT POLARITY CLASSIFICATION

Research in sentiment analysis, as we defined it, aims at classifying opinionated texts according to the polarity of the sentiment expressed. However, depending on the final use of the sentiment analysis task, classification of opinion is done at different levels.

For example, when a user is interested in finding out opinions people gave about a movie, the overall sentiment is enough to be able to decide whether to see it or not. On the other hand, when a user is interested in booking a hotel or buying an electronic product, the general opinion given in reviews might not be sufficient, because the user might be more interested in some of the features of the hotel or the electronic product and be indifferent with regard to others (e.g. location versus service in a hotel or size versus GPS capabilities in a mobile phone).

Opinion mining requires different techniques and approaches, depending both on the level of analysis that is required and interesting to the user, as well as on the type of text analyzed.

Researchers have proposed different scenarios for mining opinion and have proposed different methods for performing this task. It is important to mention that although the general idea of classifying sentiment in text is understood as one of assigning a piece of text (document, sentence, review) a value of “positive” or “negative” (or “neutral”), other scenarios were defined in which the positive category refers to “liking”, arguments brought in favor of an idea (pros) or support of a party or political view and the negative class includes expressions of “disliking” something, arguments brought against an idea expressed (cons) or opposition to an ideology.

Sentiment analysis (opinion mining) is a difficult task due to the high semantic variability of natural language, which we have defined according to the understanding given to sentiments and attitudes, supposes not only the discovery of directly expressed opinions also the extraction of phrases that indirectly or implicitly value objects, by means of emotions or attitudes.

It is also important to note the fact that sentiment analysis does not necessarily require as input an opinionated piece of text (Pang and Lee, 2008). Good versus bad news classification has also been considered as a sentiment classification task, which was approached in research such as the one proposed by Koppel and Shtrimberg (2004).

However, it is also very important to note that a clear distinction must be made at the time of performing sentiment analysis at the document level, namely that, for example, the content of good versus bad news (which is factual information) should not influence the judgment of sentiment as far as the facts are concerned or the people involved. To exemplify, a sentence such as “Great struggles have been made by the government to tackle the financial crisis, which led many companies to bankruptcy” must not be seen as negative because it discusses the consequences and gravity of the financial crisis, but must be seen as positive, when sentiment on the government is analyzed. Certainly, we can see that in this case the sentiment and its polarity arise from the manner in which the reporting is done.

We thus distinguish between document-level, sentence-level and feature-level sentiment analysis. The tasks are defined differently at each level and involve the performing of extra, more specialized steps. We will further show which those steps are. According to the survey by Pang and Lee (2008), general strategies that have been used in sentiment polarity classification were:

- Classification using the representation of text as feature vectors where entries correspond to terms, either as count of frequencies (using tf-idf), or counting the presence or absence of a certain opinion words. In this context, Wilson et al. (2005) have shown that “rare” words (that appear very infrequently in the opinion corpus), also called hapax legomena have a very good precision in subjectivity classification.

- Using information related to the part of speech of the sentiment words and applying specialized machine learning algorithms for the acquiring of such words (adjectives, verbs, nouns, adverbs). The work in acquiring nouns with sentiment has been proposed by Riloff et al. (2005). Here, the authors use dependency parsing and consider as features of machine learning algorithms the dependency relations. In this setting, information about modifiers or valence shifters can be introduced, as dependency analysis allows for the identification of the constituents that are modified.
- For the tasks in which sentiment on a certain topic must be extracted, the features used in machine learning for sentiment classifications were modified to include information on the mentions of the topic or the Named Entities mentioned in relation to it.

In the subsequent sections, we present the methods that were employed for sentiment analysis at the different levels considered.

SENTIMENT ANALYSIS AT A DOCUMENT LEVEL

Sentiment analysis has been done at a document level for movies, book reviews etc., starting from the assumption that each document (or review) focuses on a single object (product, topics) and contains opinion from a single opinion holder. This setting is true for reviews, but does not hold for newspaper articles or blog posts.

Work at this level has been done by Turney (2002), on movie reviews. The sentiment polarity of the individual opinion words is computed using a set of seed adjectives whose polarity is previously known and computing the Pointwise Mutual Information score that is obtained between the word to classify and the known word using the number of hits obtained by querying the two words together with the "NEAR" operator on the AltaVista search engine.

The final score obtained for the review is computed as sum of the polarities of the individual opinionated words in the review, from a set of sentences that is filtered according to patterns bases on the presence of adjectives and adverbs.

Another approach at the classifying polarity of sentiment at a document level is presented in Pang et al. (2002), where the authors use Naïve Bayes machine learning using unigram features and show that the use of unigrams outperforms the use of bigrams and of sentiment-bearing adjectives.

Another work in classifying documents according to their polarity is presented by Dave et al. (2003). In this work, the authors extract patterns of opinion from a corpus of reviews which are already graded.

Mullen and Collier (2004) show that classifying sentiment using Support Vector Machines with features computed on the basis of word polarity, semantic differentiation computed using synonymy patterns in WordNet, proximity to topic features and syntactic relations outperforms n-gram classifications.

Another similar approach was taken by Pang and Lee (2003). In this approach, the authors classify reviews into a larger scale of values (not only positive and negative), seen as a regression problem, and employ SVM machine learning with similarity features. They compare the outcome against the number of stars given to the review.

Chaovalit and Zhou (2005) perform a comparison between different methods of supervised and unsupervised learning based on n-gram features and semantic orientation computed by using patterns and dependency parsing.

Goldberg and Zhu (2006) present a graph-based approach to sentiment classification at a document level. They represent documents as vectors, computed on the basis of presence of opinion words and then link each document to the k most similar ones. Finally, they classify documents on the basis of the graph information using SVM machine learning.

A similar effort is made by Ng et al. (2006), where the goal is also to classify documents according to their polarity. The authors present an interesting comparison between dependency-based classification and the use of dependency relations as features for machine learning, which concludes that dependency parsing is not truly effective at the time of performing document level sentiment analysis, as it was previously shown in other research (Kudo and Matsumoto, 2004).

SENTENCE-LEVEL SENTIMENT ANALYSIS

At the sentence level, or part of document level, sentiment analysis is done in most cases in two steps: the first one views the selection of subjective sentences and the second one aims at classifying the sentiment expressed according to its polarity. The assumption that is made in this case is that each sentence expresses one single opinion.

Sentiment analysis at the sentence level includes work by Pang and Lee (2004), where an algorithm based on computing the minimum cut in a graph containing subjective sentences and their similarity scores is employed.

Yu and Hatzivassiloglou (2003) use sentence level sentiment analysis with the aim of separating fact from opinions in a question answering scenario.

Other authors use subjectivity analysis to detect sentences from which patterns can be deduced for sentiment analysis, based on a subjectivity lexicon (Hatzivassiloglou and Wiebe, 2000; Wiebe and Riloff, 2006; Wilson et al., 2004).

Kim and Hovy (2004) try to find, given a certain topic, the positive, negative and neutral sentiments expressed on it and the "source" of the opinions (the opinion holder). After creating sentiment lists using WordNet, the authors select sentences which contain both the opinion holder as well as carry opinion statements and compute the sentiment of the sentence in a window of different sizes around the target, as harmonic and, respectively, geometrical mean of the sentiment scores assigned to the opinion words.

Kudo and Matsumoto (2004) use a subtree-based boosting algorithm using dependency-tree-based features and show that this approach outperforms the bag-of-words baseline, although it does not bring significant improvement over the use of n-gram features.

FEATURE-LEVEL SENTIMENT ANALYSIS

Sentiment analysis at the feature level, also known as "feature-based opinion mining" (Hu and Liu, 2004; Liu, 2007), is defined as the task of extracting, given an "object" (product, event, person etc.), the features of the object and the opinion words used in texts in relation to the features, classify the opinion words and produce a final summary containing the percentages of positive versus negative opinions expressed on each of the features. This task has been previously defined by Dave et al. (2003).

Feature-based opinion mining involves a series of tasks:

- Task 1: Identify and extract object features that have been commented on by an opinion holder (e.g., a reviewer).
- Task 2: Determine whether the opinions on the features are positive, negative or neutral.
- Task 3: Group feature synonyms.

Subsequently, once all the groups of words referring the same feature is gathered and the polarity of the opinion is computed, the result is presented as a percentage of positive versus negative opinion on each feature (feature-based opinion summary of multiple reviews).

There are a series of techniques and approaches that were used in each of these three subtasks.

For the identification of features, "pros and cons" reviews were used, label sequential rules based on training sequences were employed to define extraction rules (Popescu and Etzioni, 2005), frequent features were mined using sequential pattern mining (frequent phrases) and patterns for "part of" relations were defined (Ding et al., 2008). Infrequent features were discovered with similarity in WordNet. Polarity classification was done using as start point the "good" and "bad" adjectives and exploring the synonyms and antonyms of these words in WordNet (Hu and Liu, 2004), using weighting functions depending on surrounding words (Popescu and

Etzioni, 2005) or using local conjunction or disjunction relations with words with priory known polarity (Ding et al., 2008). Grouping of feature synonyms was done using relations in WordNet.

An important related research area was explored in Task 18 at SemEval 2010 (Wu and Jin, 2010). In this task, the participants were given a set of contexts in Chinese, in which 14 dynamic sentiment ambiguous adjectives are selected. They were: 大|big, 小|small, 多|many, 少|few, 高|high, 低|low, 厚|thick, 薄|thin, 深|deep, 浅|shallow, 重|heavy, 轻|light, 巨大|huge, 重大|grave. The task was to automatically classify the polarity of these adjectives, i.e. to detect whether their sense in the context is positive or negative. The majority of participants employed opinion mining systems to classify the overall contexts, after which local rules were applied, depending on which the polarity surrounding the adjective to be classified remained the same as the overall polarity of the text, or it changed.

Recently, authors have shown that performing very fine or very coarse-grained sentiment analysis has drawbacks for the final application, as many times the sentiment is expressed within a context, by comparing or contrasting with it. This is what motivated McDonald et al. (2007) to propose an incremental model for sentiment analysis, starting with the analysis of text at a very fine-grained level and adding up granularity to the analysis (the inclusion of more context) up to the level of different consecutive sentences. The authors showed that this approached highly improved the sentiment analysis performance. The same observation was done by Balahur and Montoyo (2009) for the task of feature-based opinion mining and subsequently confirmed by experiments in opinion question answering (Balahur et al., 2009a; Balahur et al., 2009d; Balahur et al, 2009h; Balahur et al., 2010a; Balahur et al., 2010c).

3.2.3. EXTRACTION OF OPINION FROM TEXT

The task of extracting opinion in text refers to the problem of spotting exact parts of texts where a specifically sought opinion (a positive or negative sentiment) is presented on a specific, given “target”. The experiments conducted were used in the context of multi-perspective question answering (“Why do people like George Clooney?”), political debates (discovering what arguments are brought in favor or a against a certain law or policy), dialogues (determining who is pro or against a topic and finding the reasons) and blogs (finding users who agree or disagree with different posts and extracting their view).

The sentiment analysis tasks defined in the extraction setting generally evolve around Named Entities (the targets) and the source of the opinion is given beforehand. Interesting, state-of-the-art approaches to extracting opinion from text were implemented by systems participating in the Text Retrieval Conference (TREC) in 2006-2008, Text Analysis Conference (2008) and NTCIR MOAT competitions in 2008-2010.

State-of-the-art approaches demonstrated that the tasks related to subjectivity and sentiment analysis require the use of specialized tools, lexical resources and methods, as well as annotated corpora and gold standards against which systems can compare in order to evaluate their performance.

Different machine learning algorithms were employed in diverse settings, using several types of weighting schemes and similarity measures. Other approaches used simpler, feature vector representations of word frequencies and presence, n-gram classifications or simple bag-of-word counting of sentiment words. The use of dependency parsing was disputed as to whether is brings or not improvement over the simpler methods.

Although some approaches perform better than the others, the assumptions made when performing the different tasks make most results impossible to implement with the same success in real applications. For the latter, extensive knowledge has to be added on topic, target, source, context and even world knowledge. Last, but not least, it was proved that different textual genres and formulations of the problem of sentiment analysis require specialized approaches, either from the resource-use point of view or the need to enclose techniques from other tasks in NLP (for topic detection, IE, QA, IR).

From the overview on the state-of-the-art in the field until this point, we can conclude that we are standing in front of a challenging tasks, needing creative solutions, involving many areas of NLP and knowledge from fields outside AI or computing, such as linguistics, psychology, pragmatics, cognitive science and philosophy. Additionally, in real-case scenarios, sentiment analysis must be combined with other NLP tasks. We discuss the issues involved in this context in the next sections.

3.3. APPLICATIONS OF OPINION MINING TO TRADITIONAL TASKS – OPINION QUESTION ANSWERING, OPINION SUMMARIZATION

3.3.1. OPINION QUESTION ANSWERING

Question Answering (QA) can be defined as the Natural Language Processing (NLP) task in which given a set of questions and a collection of documents, an automatic NLP system is employed to retrieve the answer to the queries in Natural Language (NL). Research focused on building factoid QA systems has a long tradition; however, it is only recently that researchers have started to focus on the development of Opininion Question Answering (OQA) systems.

Stoyanov et al. (2005) and Pustejovsky and Wiebe (2005) studied the peculiarities of opinion questions and have found that they require the development of specific techniques to be tackled, as their answers are longer and the analysis of the question is not as strightforward as in the case of factoid questions. Cardie et al. (2004) employed opinion summarization to support a Multi-Perspective QA system, aiming to identify the opinion-oriented answers for a given set of questions.

Yu and Hatzivassiloglou (2003) separated opinions from facts and summarized them as answer to opinion questions.

Kim and Hovy (2005) identified opinion holders, which are a key component in retrieving the correct answers to opinion questions.

Due to the realized importance of blog data, recent years have also marked the beginning of NLP research focused on the development of OQA systems and the organization of international conferences encouraging the creation of effective QA systems both for fact and subjective texts. The TAC 2008[7] QA track proposed a collection of factoid and opinion queries called “rigid list” (factoid) and “squishy list” (opinion) respectively, to which the traditional QA systems had to be adapted. In this competition, some participating systems treated opinionated questions as “other” and thus they did not employ opinion specific methods. However, systems that performed better in the “squishy list” questions than in the “rigid list” implemented additional components to classify the polarity of the question and of the extracted answer snippet. The Alyssa system (Shen et al., 2007) uses a Support Vector Machines (SVM) classifier trained on the MPQA corpus (Wiebe et al., 2005), English NTCIR8 data and rules based on the subjectivity lexicon (Wilson et al., 2005). Varma et al. (2008) performed query analysis to detect the polarity of the question using defined rules. Furthermore, they filter opinion from fact retrieved snippets using a classifier based on Naïve Bayes with unigram features, assigning for each sentence a score that is a linear combination between the opinion and the polarity scores. The PolyU (Li et al., 2008b) system determines the sentiment orientation of the sentence using the Kullback-Leibler divergence measure with the two estimated language models for the positive versus negative categories. The QUANTA system (Li et al., 2008a) performs opinion question sentiment analysis by detecting the opinion holder, the object and the polarity of the opinion. It uses a semantic labeler based on PropBank[9] and manually defined patterns. Regarding the sentiment classification, they extract and classify the opinion words. Finally, for the

[7] http://www.nist.gov/tac/

[8] http://research.nii.ac.jp/ntcir/

[9] http://verbs.colorado.edu/~mpalmer/projects/ace.html

answer retrieval, they score the retrieved snippets depending on the presence of topic and opinion words and only choose as answer the top ranking results.

Other related work concerns opinion holder and target detection. NTCIR 7 MOAT organized such a task, in which most participants employed machine learning approaches using syntactic patterns learned on the MPQA corpus (Wiebe et al., 2005). Starting from the abovementioned research, the work we proposed (Balahur et al., 2009a; Balahur et al., 2009d; Balahur et al., 2009h; Balahur et al., 2010a; Balahur et al., 2010e) employed opinion specific methods focused on improving the performance of our OQA. We perform the retrieval at 1 sentence and 3 sentence-level and also determine new elements that we define as crucial for the opinion question answering scenario: the Expected Source (ES) and the Expected Target (ET), expected answer type (EAT) and Expected Polarity Type (EPT).

3.3.2. OPINION SUMMARIZATION

Whilst there is abundant literature on text summarization (Kabadjov et al., 2009; Steinberger et al., 2007; Hovy, 2005; Erkan and Radev, 2004; Gong and Liu, 2002) and sentiment analysis (Pang and Lee, 2008; Hovy et al., 2005; Kim and Hovy, 2004; Turney and Littman, 2003), there is still limited work at the intersection of these two areas (Stoyanov and Cardie, 2006; Saggion and Funk, 2010; Saggion et al., 2010).

For the first time in 2008 there was a Summarization Opinion Pilot track at the Text Analysis Conference organized by the US National Institute of Standards and Technology (NIST). The techniques employed by the participants were mainly based on the already existing summarization systems. Whilst most participants added new features (sentiment, positive/negative sentiment, positive/negative opinion) to account for the presence of positive opinions or negative ones – CLASSY (Conroy and Schlesinger, 2008); CCNU (He et al., 2008); LIPN (Bossard et al., 2008); IIITSum08 (Varma et al., 2008) -, efficient methods were proposed focusing on the retrieval and filtering stage, based on polarity - DLSIUAES (Balahur et al., 2008) - or on separating information rich clauses – italic (Cruz et al., 2008). Finally, fine-grained, feature-based opinion summarization is defined in (Hu and Liu, 2004).

The fact that opinion summarization was proposed in such a complex setting (additionally requiring the determination of answers to opinion questions), lead to very low results from the participating systems. Having realized the great challenge such an end-to-end system proposes, research in this field has been split into two directions – Opinion Question Answering and Opinion Summarization.

3.4. RELATED TASKS: EMOTION DETECTION AS A GENERAL AI AND NLP CHALLENGE

Emotion is a complex phenomenon, on which no definition that is generally accepted has been given. However, a commonly used definition considers emotion as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems (Information processing, Support, Executive, Action, Monitor) in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism". (Scherer, 1987; Scherer, 2001).

The term feeling points to a single component denoting the subjective experience process (Scherer, 2005) and is therefore only a small part of an emotion.

Moods are less specific, less intense affective phenomena, product of two dimensions - energy and tension (Thayer, 2001).

As defined by the Webster dictionary, "sentiment is a personal belief or judgment that is not founded on proof or certainty"[10].

In Artificial Intelligence (AI), the term *affective computing* was first introduced by Picard (1995). Although there were previous approaches in the 80s and 90s, in the field of NLP, the task of emotion detection has grown in importance together with the exponential increase in the volume of subjective data on the Web in blogs, forums, reviews, etc.

Previous approaches to spot affect in text include the use of models simulating human reactions according to their needs and desires (Dyer, 1987), fuzzy logic (Subasic and Huettner, 2000), lexical affinity based on similarity of contexts – the basis for the construction of WordNet Affect (Strapparava and Valitutti, 2004) or SentiWord-Net (Esuli and Sebastiani, 2005), detection of affective keywords (Riloff et al., 2003) and machine learning using term frequency (Pang et al., 2002; Wiebe and Riloff, 2006). The two latter approaches are the most widely used in emotion detection systems implemented for NLP, because they are easily adaptable across domains and languages.

Other proposed methods include the creation of syntactic patterns and rules for cause-effect modeling (Mei Lee et al., 2009).

Significantly different proposals for emotion detection in text are given in the work by (Liu et al, 2003) and the recently proposed framework of sentic computing (Cambria et al., 2009), whose scope is to model affective reaction based on commonsense knowledge.

Danisman and Alpkocak (2008) proposed an approach based on vectorial representations. The authors compute the set of words that is discriminatory for 5 of

[10] http://www.merriam-webster.com/thesaurus/

the 7 emotions in the ISEAR corpus and represent the examples using measures computed on the basis of these terms.

In the SemEval 2007 Task 18 “Affective Text” (Strapparava and Mihalcea, 2007), the task was to classify 1000 news headlines depending on their valence and emotion, using Ekman’s 6 basic emotions model (Ekman, 1999). The participating systems used rule-based or machine learning approaches, employing the polarity and emotion lexicons existent at that time (SentiWordNet, General Inquirer and WordNet Affect), the only training set available for emotion detection at that time (the training data containing 1000 news headlines provided by the task organizers) or calculating Pointwise Mutual Information scores using search engines.

In this chapter, we presented the main issues that research in sentiment analysis aims to tackle and the state-of-the-art approaches that have been proposed for them.

The methods used so far, as we have seen, do not make any distinction according to the type of text that is analyzed. In the next chapter, we present the methods and the resources that we created in ordet to tackle the task of sentiment analysis. In contrast to existing approaches, we will first study the requirements of the sentiment analysis task in the different textual genres considered, and subsequently propose adequate techniques and resources.

CHAPTER 4. METHODS AND RESOURCES FOR MULTILINGUAL SENTIMENT ANALYSIS IN DIFFERENT TEXT TYPES

Motto: *"But are not this struggle and even the mistakes one may make better and do not develop us more than if we keep systematically away from emotions?"(Vincent van Gogh)*

As we have seen in the state-of-the-art chapter, sentiment analysis can be applied to different textual genres, at a coarser or finer-grained level and for different applications. The choice in the level of analysis normally depends on the on the type of text that one is processing and the final application – i.e. on the degree of detail that one wishes or requires in order to benefit from the process of automatic sentiment detection.

While detecting the general attitude expressed in a review on a movie suffices to take the decision to see it or not, when buying an electronics product, booking a room in a hotel or travelling to a certain destination, users weigh different arguments in favor or against, depending on the "features" they are most interested in (e.g. weight versus screen size, good location versus price).

Reviews are usually structured around comments on the product characteristics and therefore, the most straightforward task that can be defined in this context is the feature-level analysis of sentiment. The feature-level analysis is also motivated by the fact that on specific e-commerce sites, reviews contain special sections where the so-called "pros" and "cons" of the products are summarized, and where "stars" can be given – to value the quality of a characteristic of a product (e.g. on a scale from 1 to 5 "stars").

As far as the source of opinion is concerned, in this type of text, reviews are written on the same topic and by the same author. At the time of processing, thus, one is not usually interested in the author of the review, but rather on being able to extract as many opinions as possible from the reviews available.

In contrast to that, in newspaper articles, for example, sentiment can be expressed on many topics within the same piece of news, by different sources. Thus, in this kind of text, the source and the target of opinions are very important at the time of analyzing opinion. Moreover, in newspaper articles, the author might convey certain opinions, by omitting or stressing upon some aspect of the text and by thus inserting their own opinion towards the facts. Such phenomena, analyzed as part of work on perspective determination or news bias research, should also be taken into consideration at the time of performing opinion mining from this textual source. Moreover, in these texts, the news in itself is highly correlated with the opinion expressed; however, the positivity or negativity of the news content should not be mistaken for the polarity of the opinion expressed therein.

In blogs, we are facing the same difficulties – i.e. of having to determine the characteristics of the source, as well as ensure that the target of the opinions expressed is the required one. Moreover, blogs have a dialogue-like structure, and most of the times, the topic discussed is related to a news item that is taken from a newspaper article. The same phenomena are also present in forums, microblogs, social network comments and reviews, but the characteristics of these texts are different (e.g. shorter documents, different language used, single versus multiple targets of opinions, different means of referencing targets).

In this chapter, we present the tasks and the methods we have proposed, in a suitable manner, to tackle sentiment analysis in different text types. For each of the textual genres considered, we have appropriately defined the task of sentiment analysis, identified the genre peculiarities and proposed adequate methods to tackle the issues found.

Where previous approaches fell short of correctly identifying the needs of the specific textual genre, we proposed adequate formulations of the problem and proposed specific methods to tackle them. Additionally, where insufficient resources were available, we have developed new annotation schemes and new corpora, for English and other languages (Spanish, German, Chinese).

This chapter is structured as follows: in Section 4.1 we present the methods and resources we proposed and evaluated in the context of sentiment analysis from product reviews. Subsequently, in Section 4.2., we discuss the issues involved in sentiment analysis from newspaper articles, specifically in reported speech extracted from news (i.e. quotations). In Section 4.3., we present a method to detect sentiments expressed in political debates, studying the needs of a generic sentiment analysis system that is able to deal with different topics, in a dialogue framework, in which different sentiment sources and targets are present. Finally, in Section 4.4., we present the methods and resources we have developed for sentiment analysis from blogs. Summing up all the experience gathered from the analysis of the previous text types, we design a general method to tackle sentiment analysis, in the context of this new and complex text type with a dialogue-like structure, in which formal and informal language styles are mixed and sentiment expressions are highly diverse.

4.1. OPINION MINING FROM PRODUCT REVIEWS – FEATURE-BASED OPINION MINING AND SUMMARIZATION

4.1.1. INTRODUCTION

Presently, the consumer market is flooded with products of the most varied sorts, each being advertised as better, cheaper, more resistant, easy to use and fault free. But is all what is advertised actually true? Certainly, all companies will claim that the products they make are the best. In practice, however, each product will behave better or worse depending on the user’s necessities and level of expertise, will have certain capabilities depending on the price and product class. Moreover, for most users, the choice will depend on reasons such as lower price, brand name or ratio between price and performance.

Therefore, how can a person that is faced with the harsh decision of having to choose among tens of products of the same type, with the same features, finally take a rational decision? Since the world wide web contains a large quantity of product reviews, in specialized review sites and user blogs, a good solution seems to be that of lending an ear to this “word-of-mouth” on the Web (Liu, 2007), weighing the pros and cons of the different products and finally buying the one satisfying the personal needs and expectations.

The clear advantage in having at hand the volume of data present on the Internet nowadays is that one is apt to obtain almost objective information on the products that he/she is planning to buy. This process can be accomplished by viewing different opinions that people who have previously bought the product in question have on it, based on the experience in using it. Such people can comment on the reasons motivating their choice of purchase and their present attitude towards that decision. Thus, besides the objective information concerning price and product capabilities, a prospect buyer can also have access to the subjective information about a product. However, a high volume of information is also bound to bring difficulty in sifting through it.

The ideal situation is that in which one is able to read all available user reviews and create his/her opinion, depending on the feature(s) of interest and the value or ratio of the feature attributes. The main problem then becomes the time spent in reviewing all available data and the language barrier the fact that product reviews are written in different languages.

The solution is a system that automatically analyzes and extracts the values of the features for a given product, independent of the language the customer review is written in. Such a system can then present the potential buyer with percentages of positive and negative opinions expressed about each of the product features and possibly make suggestions based on buyer preferences.

This can be achieved by performing sentiment analysis at the feature level, an approach that is also known as "feature-based opinion mining" (Hu and Liu, 2004; Liu, 2007). Previously mentioned by Dave et al. (2003), this task aims at extracting, given an "object" (product, event, person etc.), the features of the object and the opinions expressed in texts in relation to the features, classify the opinion words and produce a final summary containing the percentages of positive versus negative opinions expressed on each of the features. Feature-based opinion mining involves a series of tasks:

- *Task 1: Identify and extract object features that have been commented on by an opinion holder (e.g., a reviewer).*
- *Task 2: Determine whether the opinions on the features are positive, negative or neutral.*
- *Task 3: Group feature synonyms.*

Subsequently, once all the groups of words referring the same feature are gathered and the polarity of the opinion is computed, the result is presented as a percentage of positive versus negative opinion on each feature (feature-based opinion summary of multiple reviews).

The approach we use is grounded on the feature-based opinion mining and summarization paradigm, whose theoretical background has been described by Liu (2007). Relevant research in feature-driven opinion summarization has been done by Ding et al. (2008) and Dave et al. (2003).

The issues we have identified in this context is that present research has not included the discovery of implicit features and furthermore, it has left the problem of explicit features dependent on the mentioning of these features in the individual user reviews or not. The authors describe approaches that are lexicon-based and consist in discovering frequent features using association mining and determining the semantic orientation of opinions as polarity of adjectives (as opinion holders) that features are described by. The classification of adjectives is done starting with a list of seeds and completing it using the WordNet synonymy and antonymy relations. Infrequent features are deduced using the opinion holders. However, the fact that there is no well-organized structure of features and sub-features of products leads to the fact that, for example, the summarization of opinions is done for 720 features for an mp3 player (Ding et al., 2008). The question that arises is: would a user in a real-life situation be interested on whether the edges of a camera are round or flat and what the previous buyers think about that, or would a potential buyer like to see if the design of the product is fine or not, according to the many criteria developed by buyers to assess this feature? The work does not approach implicit features and does not classify the orientation of adjectives depending on the context. A solution to the latter problem is presented by Ding et al. (2008) where the authors take a holistic approach to classifying adjectives, that is, consider not only the local context in which they appear next to the feature they determine, but also other adjectives appearing with the feature and their polarity in different contexts. Popescu and Etzioni (2005) employ a more complex approach for feature-based summarization of opinions, by computing the web PMI (Pointwise Mutual Information) statistics for the explicit feature extraction and a technique called relaxation labeling for the assignation of polarity to the opinions. In this approach, dependency parsing is used together with ten extraction rules that were developed intuitively.

We propose an initial approach to the issue of feature-based opinion mining (Balahur and Montoyo, 2008d; Balahur and Montoyo, 2008f), which we subsequently extended in Balahur and Montoyo (2008c), by introducing product technical details and comparing two different measures of term relatedness (Normalized Google Distance (Cilibrasi and Vitanyi, 2006) and Latent Semantic Analysis (Deerwester et al., 1990). On top of this system, we propose a method to recommend products based on the scores obtained for the different quantified features in the opinion mining step. The method was presented in (Balahur and Montoyo, 2008b).

In the light of the fact that no standard annotation was available for feature-based opinion mining, in Balahur and Montoyo (2009) we proposed an annotation scheme that aimed at standardizing the labeling of reviews, so that different types of mentions of features (direct, indirect, implicit) and the different manner of expressing opinions (through subjective or objective statements) can be correctly labeled. Subsequently, we studied methods to infer sentiment expressed on different features using examples of annotations from a review corpus and Textual Entailment (Balahur and Montoyo, 2009).

In the following sections we describe our approaches and results. The method we propose is language and customer-review independent. It extracts a set of general product features, finds product specific features and feature attributes and is thus applicable to all possible reviews in a product class. The approaches we present in this section were summarized by Balahur et al. (2010).

4.1.2. METHODS FOR MINING SENTIMENT FROM PRODUCT REVIEWS

Our method consists of two distinct steps: preprocessing and main processing, each containing a series of sub-modules and using different language tools and resources.

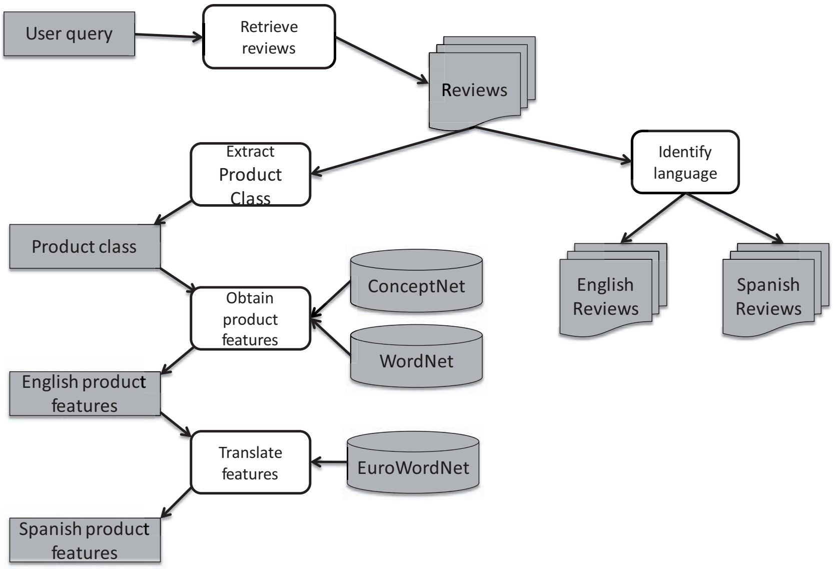


Figure 4.1: Preprocessing Stage

A PREPROCESSING

In our approach, we start from the following scenario: a user enters a query about a product that he/she is interested to buy.

The search engine retrieves a series of documents containing the product name, in different languages. Further on, two parallel operations are performed: the first one uses the Lextek[11] language identifier software to filter and obtain two categories one containing the reviews in English and the other the reviews in Spanish.

The second operation implies a modified version of the system proposed by Kozareva et al. (2007) for the classification of person names. We use this system in order to determine the category that the product queried belongs to (e.g. digital camera, laptop, printer, book). Once the product category is determined, we proceed to extracting the product specific features and feature attributes. This is accomplished using WordNet and ConceptNet and the corresponding mapping to Spanish using EuroWordNet. Apart from the product specific class of features and feature attributes, we consider a core of features and feature attributes that are product-independent and whose importance determines their frequent occurrence in customer reviews. Figure 4.1 describes the components used in the preprocessing stage.

[11] http://www.lextek.com/langid/

Product-independent features and feature attributes

There are a series of features that are product independent and that are important to any prospective buyer. We consider these as forming a core of product features. For each of these concepts, we retrieve from WordNet the synonyms which have the same Relevant Domain (Vázquez et al., 2004), the hyponyms of the concepts and their synonyms and attributes, respectively.

Using WordNet to extract product specific features and feature attributes

Once the product category has been identified, we use WordNet to extract the product specific features and feature attributes. We accomplish this in the following steps:

1. For the term defining the product category, we search its synonyms in WordNet (Fellbaum , 1999).
2. We eliminate the synonyms that do not have the same top relevant domain (Vázquez et al., 2004) as the term defining the product category.
3. For the term defining the product, as well as each for each of the remaining synonyms, we obtain their meronyms from in WordNet, which constitute the parts forming the product.
4. Since WordNet does not contain much detail on the components of most of new technological products, we use ConceptNet to complete the process of determining the specific product features. We explain the manner in which we use ConceptNet in the following section. After performing the steps described above, we conclude the process of obtaining the possible terms that a customer buying a product will comment on. The final step consists in finding the attributes of the features discovered by applying the *"has attributes"* relation in WordNet to each of the nouns representing product features. In the case of nouns which have no term associated by the has attribute relation, we add as attribute features the concepts found in ConceptNet under the *OUT* relations *PropertyOf* and *CapableOf*. In case the concepts added are adjectives, we further add their synonyms and antonyms from WordNet.

As result we have for example, in the case of "photo", the parts "raster" and "pixel" with the attributes "blurry", "clarity", "sharp".

Using ConceptNet to extract product-specific features and feature attributes

ConceptNet (Liu and Singh, 2004) is a freely available commonsense knowledgebase and NLP toolkit which supports many practical textual reasoning tasks over real-world documents. Commonsense knowledge in ConceptNet encompasses the spatial, physical, social, temporal, and psychological aspects of everyday life. It contains relations such as *CapableOf, ConceptuallyRelatedTo, IsA, LocationOf* etc. In order to obtain additional features for the product in question, we add the concepts that are related to the term representing the concept with terms related in ConceptNet by the *OUT* relations *UsedFor* and *CapableOf* and the *IN* relations *PartOf* and *UsedFor*. For example, for the product "camera", the *OUT UsedFor* and *CapableOf* relations that will added are "take picture", "take photograph", "photography", "create image", "record image" and for the *IN PartOf* and *UsedFor* relations "shutter", "viewfinder", "flash", "tripod".

Mapping concepts using EuroWordNet

EuroWordNet[12] (EWN) is a multilingual database with WordNets for different European languages (Dutch, Italian, Spanish, German, French, Czeck and Estonian). Each language has its own designed WordNet, structured as the Princeton WordNet. Having these connections, it is possible that parting from one word, one can consult similar words in any other language of the EWN. The main advantage in using this lexical resource is that all the terms discovered in one language can be easily mapped to another language. We employ EuroWordNet and map the features and feature attributes, both from the main core of words, as well as the product specific ones that were previously discovered for English, independent of the sense number, taking into account only the preservation of the relevant domain. Certainly, we are aware of the noise introduced by this mapping, however in the preliminary research we found that the concepts introduced that had no relation to the product queried did not appear in the user product reviews.

Discovering overlooked product features

The majority of product features we have identified so far are parts constituting products. However, there remains a class of undiscovered features that are indirectly related to the product. These are the features of the product constituting parts, such as *"battery life"*, *"picture resolution"*, *"auto mode"*. Further, we propose to extract these overlooked product features by determining bigrams made up of target words constituting features and other words in a corpus of customer reviews. In the case of digital cameras, for example, we considered a corpus of 200 customer reviews on which we ran Pedersen's Ngram Statistics Package to determine target co-occurrences of the features identified so far. As measure for term association, we use the Pointwise Mutual Information (PMI) score, which is calculated according to the following formula:

$$PMI(x,y) = \frac{P(x,y)}{P(x)P(y)}$$

[12] http://www.illc.uva.nl/EuroWordNet/

where x and y are two words and P(x) stands for the probability of the word x occurring in the corpus considered. In this manner, we discover bigram features such as “battery life”, “mode settings” and “screen resolution”.

B MAIN PROCESSING

The main processing in our system is done in parallel for English and Spanish. In the next section, we will briefly describe the steps followed in processing the initial input containing the customer reviews in the two considered language and offer as output the summarized opinions on the features considered. Figure 4.2 presents the steps included in the processing.

We start from the reviews filtered according to language. For each of the two language considered, we used a specialized tool for anaphora resolution- JavaRAP for English and SUPAR (Ferrández et al., 1999) for Spanish. Further on, we separate the text into sentences and use a Named Entity Recognizer to spot names of products, brands or shops.

Using the lists of general features and feature attributes, product-specific features and feature attributes, we extract from the set of sentences contained in the text only those containing at least one of the terms found in the lists.

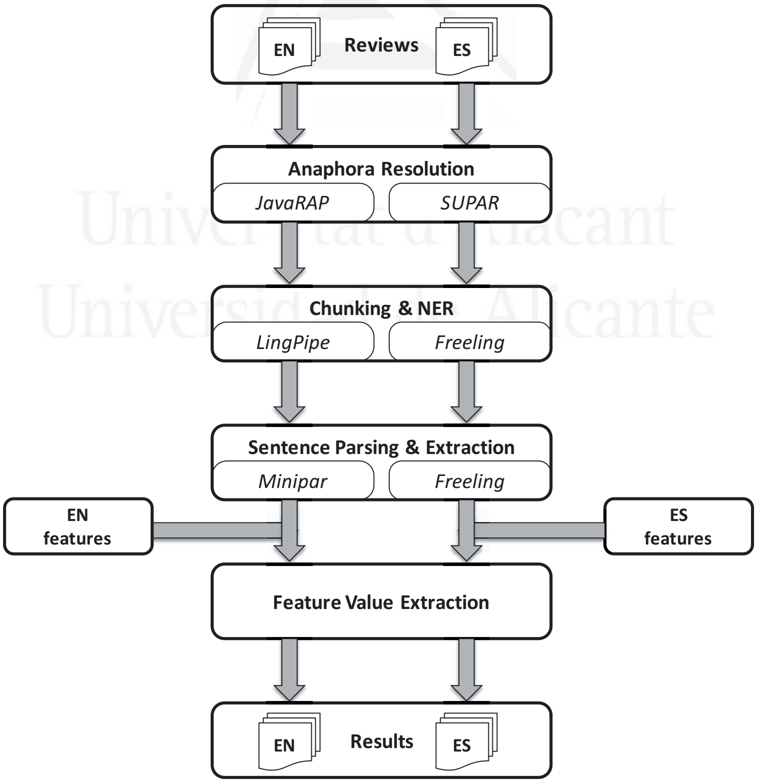


Figure 4.2: System Architecture

Anaphora resolution

In order to solve the anaphoric references on the product features and feature attributes, we employ two anaphora resolution tools - JavaRAP for English and SUPAR for Spanish. Using these tools, we replace the anaphoric references with their corresponding referents and obtain a text in which the terms constituting product features could be found.

JavaRAP is an implementation of the classic Resolution of Anaphora Procedure (RAP) given by Lappin and Leass (1994). It resolves third person pronouns, lexical anaphors, and identifies pleonastic pronouns.

Using JavaRAP, we obtain a version of the text in which pronouns and lexical references are resolved. For example, the text: *"I bought this camera about a week ago, and so far have found it very very simple to use, takes good quality pics for what I use it for (outings with friends/family, special events). It is great that it already comes w/ a rechargeable battery that seems to last quite a while..."*, by resolving the anaphoric pronominal reference, becomes *"I bought this camera about a week ago, and so far have found* <this camera> *very very simple to use, takes good quality pics for what I use* <this camera> *for (outings with friends/family, special events). It is great that* <this camera> *already comes w/a rechargeable battery that seems to last quite a while..."*.

For the anaphora resolution in Spanish, we employ SUPAR (Slot Unification Parser for Anaphora Resolution). The architecture of SUPAR contains, among others, a module solving the linguistic problems (pronoun anaphora, element extraposition, ellipsis, etc.). We use SUPAR in the same manner as JavaRAP, to solve the anaphora for Spanish. Sentence chunking and NER Further on, we split the text of the customer review into sentences and identify the named entities in the text. Splitting the text into sentences prevents us from processing sentences that have no importance as far as product features that a possible customer could be interested in are concerned.

Chunking and Named Entity Resolution

LingPipe[13] is a suite of Java libraries for the linguistic analysis of human language. It includes features such as tracking mentions of entities (e.g. people or proteins), part-of-speech tagging and phrase chunking.

We use LingPipe to split the customer reviews in English into sentences and identify the named entities referring to products of the same category as the product queried. In this manner, we can be sure that we identify sentences referring to the

[13] http://alias-i.com/lingpipe/

product queried, even the reference is done by making use of the name of another product. For example, in the text "For a little less, I could have bought the Nikon Coolpix, but it is worth the extra money.", anaphora resolution replaces `<it>` with `<Nikon Coolpix>` and this step will replace it with `<camera>`.

The FreeLing[14] package consists of a library providing language analysis services. The package offers many services, among which text tokenization, sentence splitting, POS-tagging, WordNet based sense annotation and rule-based dependency parsing. We employ FreeLing in order to split the customer reviews in Spanish into sentences and identify the named entities referring to products of the same category as the product queried.

Sentence extraction

Having completed the feature and feature attributes identification phase, we proceed to extracting for further processing only the sentences that contain the terms referring to the product, product features or feature attributes. In this manner, we avoid further processing of text that is of no importance to the task we wish to accomplish. For example, sentences of the type *"I work in the home appliances sector"* will not be taken into account in further processing. Certainly, at the overall level of review impact, such a sentence might be of great importance to a reader, since it proves the expertise of the opinion given in the review. However, for the problems we wish to solve by using this method, such a sentence is of no importance.

Sentence parsing

Each of the sentences that are filtered by the previous step are parsed in order to obtain the sentence structure and component dependencies. In order to accomplish this, we use Minipar (Lin, 1998) for English and FreeLing for Spanish. This step is necessary in order to be able to extract the values of the features mentioned based on the dependency between the attributes identified and the feature they determine.

Feature value extraction

Further on, we extract features and feature attributes from each of the identified sentences, using the following rules:

1. We introduce the following categories of context polarity shifters (Polanyi and Zaenen, 2004), in which we split the modifiers and modal operators in two categories – i.e. positive and negative:
    - *negation:* no, not, never, etc.

[14] http://nlp.lsi.upc.edu/freeling/

   - *modifiers:* positive (extremely, very, totally, etc.) and negative (hardly, less, possibly, etc.)
   - *modal operators:* positive (must, has) and negative (if, would, could, etc.)
2. For each identified feature that is found in a sentence, we search for a corresponding feature attribute that determines it. Further on, we search to see if the feature attribute is determined by any of the defined modifiers. We consider a variable we name *valueOfModifier*, with a default value of -1, that will account for the existence of a positive or negative modifier of the feature attribute. In the affirmative case, we assign a value of 1 if the modifier is positive and a value of 0 if the modifier is negative. If no modifier exists, we consider the default value of the variable. We extract triplets of the form *(feature, attributeFeature, valueOfModifier).* In order to accomplish this, we use the syntactic dependency structure of the phrase, we determine all attribute features that determine the given feature (in the case of Minipar, they are the ones connected by the *"det"* or *"mod"* relation).
3. If a feature attribute is found without determining a feature, we consider it to implicitly evoke the feature that it is associated with in the feature collection previously built for the product. *"The camera is small and sleek."* becomes *(camera, small, -1)* and *(camera, sleek, -1)*, which is then transformed by assigning the value "small" to the "size" feature and the value "sleek" to the "design" feature.

Assignment of Polarity to Feature Attributes

In order to assign polarity to each of the identified feature attributes of a product, we employ Support Vector Machines Sequential Minimal Optimization (SVM SMO) machine learning (Platt, 1998) and the Normalized Google Distance (NGD).

The main advantage in using this type of polarity assignment is that NGD is language independent and offers a measure of semantic similarity taking into account the meaning given to words in all texts indexed by Google from the world wide web. The set of anchors contains the terms *{featureName, happy, unsatisfied, nice, small, buy}*, that have possible connection to all possible classes of products.

Further on, we build the classes of positive and negative examples for each of the feature attributes considered. From the corpus of annotated customer reviews, we consider all positive and negative terms associated to the considered attribute features. We then complete the lists of positive and negative terms with their WordNet synonyms. Since the number of positive and negative examples must be equal, we will consider from each of the categories a number of elements equal to the size of the smallest set among the two, with a size of at least 10 and less or equal with 20. We give as example the classification of the feature attribute “tiny”, for the “size” feature. The set of positive feature attributes considered contains 15 terms (e.g. big, broad, bulky, massive, voluminous, large-scale, etc.) and the set of negative feature attributes considered is composed as opposed examples, such as (small, petite, pocket-sized, little, etc).

We use the anchor words to convert each of the 30 training words to 6-dimensional training vectors defined as v(j,i) = NGD(w_i,a_j), where a_j with j ranging from 1 to 6 are the anchors and w_i, with i from 1 to 30 are the words from the positive and negative categories.

After obtaining the total 180 values for the vectors, we use SVM SMO to learn to distinguish the product specific nuances. For each of the new feature attributes we wish to classify, we calculate a new value of the vector vNew(j,word)=NGD(word, a_j), with j ranging from 1 to 6 and classify it using the same anchors and trained SVM model.

In the example considered, we had the following results (we specify between brackets the word to which the scores refer to:

(small)1.52,1.87,0.82,1.75,1.92,1.93,positive
(little)1.44,1.84,0.80,1.64,2.11,1.85,positive
(big)2.27,1.19,0.86,1.55,1.16,1.77,negative
(bulky)1.33,1.17,0.92,1.13,1.12,1.16,negative

The vector corresponding to the “tiny” attribute feature is:

(tiny)1.51,1.41,0.82,1.32,1.60,1.36.

This vector was classified by SVM as positive, using the training set specified above. The precision value in the classifications we made was between 0.72 and 0.80, with a kappa value above 0.45.

For each of the features identified, we compute its polarity depending on the polarity of the feature attribute that it is determined by and the polarity of the context modifier the feature attribute is determined by, in case such a modifier exists. Finally, we statistically summarize the polarity of the feature attributes, as ratio between the number of positive quantifications and the total number of quantifications made in the considered reviews to that specific feature and as ratio between the number of negative quantifications and the total number of quantifications made in all processed reviews. The formulas can be summarized in:

1. $F_{\text{pos}}(i) = \frac{\#\text{pos feature attributes}(i)}{\#\text{feature atributes}(i)}$

2. $F_{\text{neg}}(i) = \frac{\#\text{neg feature attributes}(i)}{\#\text{feature atributes}(i)}$

The results shown are triplets of the form (feature, percentagePositiveOpinions, percentageNegativeOpinions).

C. DISCUSSION AND EVALUATION

For the evaluation of the system, we annotated a corpus of 50 customer reviews for each language, collected from sites as amazon.com, newegg.com, dealsdirect.com, ciao.es, shopmania.es, testfreaks.es and quesabesde.com. The corpus was annotated at the level of feature attributes, by the following scheme:

```
<attribute>(name of attribute)

<feature>(feature it determines)</feature>

<value>(positive/ negative)</value>

</attribute>
```

It is difficult to evaluate the performance of such a system, since we must take into consideration both the accuracy in extracting the features that reviews comment on, as well as the correct assignation of identified feature attributes to the positive or negative category.

The formula used in measuring the accuracy of the system represented the normalized sum of the ratios between the number of identified positive feature attributes and the number of existing positive attributes and the ratio of identified negative feature and the total number of negative feature attributes for each of the considered features existing in the text.

Secondly, we compute the Feature Identification Precision (P) as ratio between the number of features correctly identified from the features identified and the number of identified features.

Thirdly, we compute the Feature Identification Recall (R) as the number of correctly identified features from the features identified and the number of correctly identified features. The results obtained are summarized in Table 4.1.

We show the scores for each of the two languages considered separately and the combined score when using both systems for assigning polarity to feature attributes of a product. In the last column, we present a baseline, calculated as average of using the same formulas, but taking into consideration, for each feature, only the feature attributes we considered as training examples for our method.

Feature extraction performance	English	Spanish	Combined	Baseline English	Baseline Spanish
Accuracy	0.82	0.80	0.81	0.21	0.19
Precision	0.80	0.78	0.79	0.20	0.20
Recall	0.79	0.79	0.79	0.40	0.40

Table 4.1: System results on the annotated review corpus

We can notice how the use of NGD helped the system acquire significant new knowledge about the polarity of feature attributes, in the context.

There are many aspects to be taken into consideration when evaluating a system identifying features, opinion on features and summarizing the polarity of features. First of all, customers reviewing products on the web frequently use informal language, disregard spelling rules and punctuation marks.

At times, phrases are pure enumerations of terms, containing no subject or predicate. In this case, when there is no detectable dependency structure between components, an alternative method should be employed, such as verifying if the terms appearing near the feature within a window of specified size are frequently used in other contexts with relation to the feature. Secondly, there are many issues regarding the accuracy of each of the tools and language resources employed and a certain probability of error in each of the methods used. In this initial research, we presented a method to extract, for a given product, the features that could be commented upon in a customer review.

Further, we have shown a method to acquire the feature attributes on which a customer can comment in a review. Moreover, we presented a method to extract and assign polarity to these product features and statistically summarize the polarity they are given in the review texts in English and Spanish. The method for polarity assignment is largely language independent (it only requires the use of a small number of training examples) and the entire system can be implemented in any language for which similar resources and tools as the ones used for the presented system exist.

The main advantage obtained by using this method is that one is able to extract and correctly classify the polarity of feature attributes, in a product dependent manner. Furthermore, the features in texts are that are identified are correct and the percentage of identification is high. Not lastly, we employ a measure of word similarity that is in itself based on the "word-of-mouth" on the web. The main disadvantage consists in the fact that SVM learning and classification is dependent on the NGD scores obtained with a set of anchors that must previously be established. This remains a rather subjective matter. Also, the polarity given in the training set determines the polarity given to new terms, such that “large” in the context of “display” will be trained as positive and in the case of “size” as negative. However, there are many issues that must be addressed in systems identifying customer opinions on different products on the web. The most important one is that concerning the informal language style, which makes the identification of words and dependencies in phrases sometimes impossible.

4.1.3. CLASSIFYING OPINION ON PRODUCTS USING RELATEDNESS SCORES ON SPECIFIC CORPORA AND EMOTION EXPRESSION PATTERNS

Subsequently to this first approach, we improved the system by adding extra features, taking into consideration the product technical specifications and defining patterns for indirectly expressed opinions using WordNet Affect categories (Balahur and Montoyo, 2008c), as well as enriching our feature-dependent method of opinion classification using Latent Semantic Analysis relatedness scores; in the case of LSA scores, the context is given by the corpus from which the model is learnt, as opposed to the NGD, which is computed at the web level. We show the manner in which all these factors influence the system performance and at what cost. Last, but not least, many of the opinions on products are expressed in an indirect manner, that is, not relating the product or its features with polarity words, but expressing an emotion about them. We propose a set of patterns to extract such indirectly expressed opinions using the emotion lists from WordNet Affect.

Our solution to the problem of feature attributes classification is using machine learning with two measures of similarity. On the one hand, we employ the Normalized Google Distance, which gives a measure of the strength of relationship between two considered words at the level of the entire Web and on the other hand, we use the LSA, which gives the same measure of strength, but at a local corpus level. Classifying the feature attributes according to these scores and taking into consideration 6 anchor words that relate each word with the feature and known polarities, we show how the classification of feature attributes can be done in the feature context.

In the reviews to be mined and summarized, however, other opinion words can be found and other manners of expressing opinion can be encountered, such as those describing emotional states related to the product (e.g. *“I love this camera”*) or to using it. The solution we give to this problem is to propose a list of patterns to extract from the reviews such phrases containing emotion words, which are used to express opinions of the different product features using descriptions of emotions felt. The words related to emotion are taken from WordNet Affect.

In the evaluation section, we show how the use of such patterns raised with 12% the recall of the system, while the precision of classification rose to the same degree. In our previous approach, in order to assign polarity to each of the identified feature attributes of a product, we employed SVM SMO machine learning and the NGD. In this approach, we complete the solution with a classification employing Latent Semantic Analysis with Support Vector Machines classification. For this, we build the classes of positive and negative examples for each of the feature attributes considered. From the list of classified feature attributes in the pros and cons reviews, we consider all positive and negative terms associated to the considered attribute features. We then complete the lists of positive and negative terms with their WordNet synonyms. Since the number of positive and negative examples must be equal, we will consider from each of the categories a number of elements equal to the size of the smallest set among the two, with a size of at least 10 and less or equal with 20.

We give as example the classification of the feature attribute "tiny", for the "size" feature. The set of positive feature attributes considered contains 15 terms such as "big", "broad", "bulky", "massive", "voluminous", "large-scale" etc. and the set of negative feature attributes considered is composed as opposed examples, such as "small", "petite", "pocket-sized", "little" etc. We use the anchor words to convert each of the 30 training words to 6-dimensional training vectors defined as v(j,i) = LSA(w_i,a_j), where a_j with j ranging from 1 to 6 are the anchors and w_i, with i from 1 to 30 are the words from the positive and negative categories. After obtaining the total 180 values for the vectors, we use SVM SMO to learn to distinguish the product specific nuances. For each of the new feature attributes we wish to classify, we calculate a new value of the vector vNew(j,word) = LSA(word, a_j), with j ranging from 1 to 6 and classify it using the same anchors and trained SVM model. We employed the classification on the corpus present for training in the Infomap software pack. The blank lines represent the words which were not found in the corpus; therefore a LSA score could not be computed. The results are presented in Table 4.2. On the other hand, we employed the classification on a corpus made up of reviews on different electronic products, gathered using the Google API and a site restriction on "amazon.com". In the table below, we show an example of the scores obtained with LSA on the features attributes classified for the feature "size". The vector for the feature attribute "tiny" was classified by SVM as positive, using the training set specified above. The results are presented in Table 4.3.

Feature attribute	V1	V2	V3	V4	V5	V6	Polarity
small	0.76	0.74	---	0.71	1	0.71	pos
big	0.80	0.75	---	0.74	0.73	0.68	neg
bulky	---	---	---	---	---	---	pos
little	---	---	---	---	---	---	neg

Feature attribute	V1	V2	V3	V4	V5	V6	Polarity
tiny	0.81	0.71	---	0.80	0.73	0.72	---

Table 4.2: LSA scores on non-specialized corpus (not only with product reviews)

In Table 4.3, we show an example of the scores obtained with the similarity given by the LSA scores on a specialized corpus of reviews on products. The vector for the feature attribute "tiny" was classified by SVM as positive, using the training set specified above.

Feature attribute	V1	V2	V3	V4	V5	V6	Polarity
small	0.83	0.77	0.48	0.72	1	0.64	pos
big	0.79	0.68	0.74	0.73	0.77	0.71	neg
bulky	0.76	0.67	0.71	0.75	0.63	0.78	pos
little	0.82	0.76	0.52	0.71	0.83	0.63	neg
tiny	0.70	0.70	0.65	0.67	0.71	0.71	pos

Table 4.3: LSA scores on a specialized corpus of product reviews

Precision values in classifications we made with NGD and LSA for different product features for the examples of digital camera reviews and the mobile phones reviews vary from 0.75 to 0.8 and kappa statistics shows high confidence of classification (Balahur and Montoyo, 2008c).

The conclusion that can be drawn from the results presented is that the main advantage in using the first method of polarity assignment is that NGD is language independent and offers a measure of semantic similarity taking into account the meaning given to words in all texts indexed by Google from the World Wide Web.

On the other hand, using the whole Web corpus can also add significant noise. Therefore, we employ Latent Semantic Analysis at a local level, both on a non-specialized corpus, as well as on a corpus containing customer reviews. As we will show, the classification using LSA on a specialized corpus brings an average of 8% of improvement in the classification of polarity and a rise of 0.20 in the kappa measure, leading to an 8% overall improvement in the precision of the summarization system. However, these results were obtained using a specialized corpus of opinions, which was previously gathered from the Web. To this respect, it is important to determine sources (web sites, blogs or forums) specific to each of the working languages, from which to gather the corpus on which the LSA model can be built.

Using LSA on a non-specialized corpus improved the classification to the same degree as the classification on a specialized corpus in the cases where the specific pairs of words to be classified were found in the corpus. However, in 41% of the cases, the classification failed due to the fact that the words we tried to classify were not found in the corpus. Further on, we developed a method for feature polarity extraction using subjective phrases.

As observed before, some opinions on the product or its features are expressed indirectly, with subjective phrases containing positive or negative emotions which are related to the product name, product brand or its features. In order to identify those phrases, we have constructed a set of rules for extraction, using the emotion lists from WordNet Affect. For the words present in the “joy” emotion list, we consider the phrases extracted as having a positive opinion on the product or the feature contained. For the words in the “anger”, “sadness” and “disgust” emotion lists, we consider the phrases extracted as having a negative opinion on the product or the feature contained. Apart from the emotion words, we have considered a list of “positive words” (pos list), containing adverbs such as “definitely”, “totally”, “very”, “absolutely” and so on - as words positively stressing upon an idea - (Iftene and Balahur-Dobrescu, 2007), that influence on the polarity of the emotion expressed and that are often found in user reviews.

We present the extraction rules in Table 4.4 (verb emotion, noun emotion and adj emotion correspond to the verbs, nouns and adjectives, respectively, found in the emotion lists from WordNet Affect under the emotions “joy”, “sadness”, “anger” and “disgust”). In case of “surprise”, as emotion expressed about a product and its features, it can have both a positive, as well as negative connotation. Therefore, we have chosen not to include the terms expressing this emotion in the extraction patterns.

1. I [pos list*][verb emotion][this||the||my] [product name||product feature]
2. I ([am||’m||was||feel||felt])([pos list**])[adj emotion][with||about||by] [product name||product feature]
3. I [feel||felt][noun emotion][about||with][product name ||product brand]
4. I [pos list*][recommend][this||the][product name||product brand]
5. I ([don’t])[think ||believe][sentence***]
6. It [’s||is] [adj emotion] [how||what][product name||product feature][product action]
7. You ||Everybody||Everyone||All||He||She||They][will||would][verb emotion][this||the][product name brand||feature]

Table 4.4: List of patterns for opinion extraction based on emotion clues

We have performed a comparative analysis of the system employing the SVM SMO polarity classification using NGD and LSA on a specialized corpus, the subjective phrases and combined, with the corpus used by Balahur and Montoyo (Balahur and Montoyo, 2008c) and also the corpus of 5 reviews from (Hu and Liu, 2004). Results obtained in Table 4.5 are obtained when evaluating on our own annotated corpus, in terms of Precision (P) and Recall (R).

NGD		**LSA**		**Rules**		**NGD+ Rules**		**LSA+ Rules**	
P	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0.80	0.79	0.88	0.87	0.32	0.6	0.89	0.85	0.93	0.93

Table 4.5: System results on the review test set in Balahur and Montoyo(2008c)

In the case of the 5-review corpus proposed by Hu and Liu (2004), the observation that is important to make is that, as opposed to the annotation made in the corpus, we have first mapped the features identified to the general feature of the product (for example “fit” refers to “size” and “edges” refers to “design”), as we relieve that in real life situations, a user benefits more from a summary on coarser classes of product features.

NGD		**LSA**		**Rules**		**NGD+ Rules**		**LSA+ Rules**	
P	**R**	**P**	**R**	**P**	**P**	**R**	**P**	**R**	**P**
0.81	0.80	0.85	0.88	0.28	0.5	0.89	0.85	0.93	0.93

Table 4.6: System results on the corpus employed by Hu and Liu (2004)

Also, a set of sentences that were not annotated in the corpus, such as “You’ll love this camera”, which expresses a positive opinion on the product. The results shown in Table 4.6 are compared against the baseline of 0.20 precision and 0.41 recall, which was obtained using only the features determined by Balahur and Montoyo (Balahur and Montoyo, 2008f) and the feature attributes whose polarity was computed from the “pros and cons”-style reviews. As it can be seen, the best results are obtained when using the combination of LSA with the rules for subjective phrases extraction. However, gathering the corpus for the LSA model can be a costly process, whereas NGD scores are straightforward to be obtained and classifying is less costly as time and resources used.

What is interesting to study is the impact of employing LSA for gradual learning and correction of a system that uses NGD for classifying the polarity of feature attributes. In such a self-learning scheme, the "online" classification would be that of NGD. However, the classification of the new feature attributes can be later improved "offline" using the classification given by LSA, which can then be used as better training for learning the polarity of new feature attributes by the "online" NGD classification.

From this subsequent research, we could draw some conclusions on the advantages and disadvantages of using different scenarios for computing opinion polarity. The main advantage in using polarity assignment depending on NGD scores is that this is language independent and offers a measure of semantic similarity taking into account the meaning given to words in all texts indexed by Google from the World Wide Web. The main advantage in using LSA on a specialized corpus, on the other hand, is that it eliminates the noise given by the multiple senses of words. We completed the opinion extraction on different product features with rules using the words present in WordNet Affect, as indicative of indirectly expressed opinions on products. We showed how all the employed methods led to significant growth in the precision and recall of our opinion mining and summarization system.

4.1.4. FURTHER IMPROVEMENTS TO SENTIMENT ANALYSIS IN REVIEWS – THE ISSUE OF "DISAMBIGUATING ADJECTIVE POLARITY" USING THE LOCAL CONTEXT

As we have seen in the previous section, one of the challenges faced in opinion mining is the fact that some adjectives have a different polarity depending on the context in which they appear.

Our initial approach consisted in classifying adjective polarity using a set of anchors and the NGD and LSA scores, respectively. However, this component was not evaluated separately. With the participation in the SemEval-2 Task Number 18 – Disambiguation of Sentiment Ambiguous Adjectives (Wu and Jin, 2010), we aimed at extending the feature-based opinion mining system to resolve this task in a more general opinion mining scenario. Thus, we aimed at proposing a suitable method to tackle this issue, in a manner that is independent from the feature-based opinion mining framework. We named the system participating in this competition OpAL (Balahur and Montoyo, 2010).

In this task, the participants were given a set of contexts in Chinese, in which 14 dynamic sentiment ambiguous adjectives are selected. They were: 大|big, 小|small, 多|many, 少|few, 高|high, 低|low, 厚|thick, 薄|thin, 深|deep, 浅|shallow, 重|heavy, 轻|light, 巨大|huge, 重大|grave. The task was to automatically classify the polarity of these adjectives, i.e. to detect whether their sense in the context is positive or negative.

Our approach is based on three different strategies: a) the evaluation of the polarity of the whole context using an opinion mining system; b) the assessment of the polarity of the local context, given by the combinations between the closest nouns and the adjective to be classified; c) rules aiming at refining the local semantics through the spotting of modifiers. The final decision for classification is taken according to the output of the majority of these three approaches. The method used yielded good results, the OpAL system run achieving approximately 76% micro accuracy on a Chinese corpus. In the following subsections, we explain more in detail the individual components employed.

A. THE OPAL OPINION MINING COMPONENT

First, we process each context using Minipar[15]. We compute, for each word in a sentence, a series of features, computed from the NTCIR 7 data and the EmotiBlog annotations. These words are used to compute vectors of features for each of the individual contexts:

- the part of speech (POS)
- opinionatedness/intensity - if the word is annotated as opinion word, its polarity, i.e. 1 and -1 if the word is positive or negative, respectively and 0 if it is not an opinion word, its intensity (1, 2 or 3) and 0 if it is not a subjective word
- syntactic relatedness with other opinion word – if it is directly dependent of an opinion word or modifier (0 or 1), plus the polarity/intensity and emotion of this word (0 for all the components otherwise)
- role in 2-word, 3-word, 4-word and sentence annotations: opinionatedness, intensity and emotion of the other words contained in the annotation, direct dependency relations with them if they exist and 0 otherwise.

We add to the opinion words annotated in EmotiBlog the list of opinion words found in the Opinion Finder, Opinion Finder, MicroWordNet Opinion, General Inquirer, WordNet Affect, emotion triggers lexical resources. We train the model using the SVM SMO implementation in Weka[16].

[15] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm
[16] http://www.cs.waikato.ac.nz/ml/weka/

B. ASSESSING LOCAL POLARITY USING GOOGLE QUERIES

This approach aimed at determining the polarity of the context immediately surrounding the adjective to be classified. To that aim, we constructed queries using the noun found before the adjective in the context given, and issued six different queries on Google, together with six pre-defined adjectives whose polarity is known (3 positive - “positive”, “beautiful”, “good” and 3 negative – “negative”, “ugly”, “bad”). The form of the queries was “noun+adjective+AND+pre-defined adjective”. The local polarity was considered as the one for which the query issued the highest number of total results (total number of results for the 3 queries corresponding to the positive adjectives or to the negative adjectives, respectively).

C. MODIFIER RULES FOR CONTEXTUAL POLARITY

This rule accounts for the original, most frequently used polarity of the given adjectives (e.g. *high* is *positive, low* is *negative)*. For each of them, we define its default polarity. Subsequently, we determine whether in the window of 4 words around the adjective there are any modifiers (valence shifters). If this is the case, and they have an opposite value of polarity, the adjective is assigned a polarity value opposite from its default one (e.g. *too high* is *negative)*. We employ a list of 82 positive and 87 negative valence shifters.

D. EVALUATION RESULTS FOR OPAL IN THE SEMEVAL-2010 TASK 18

Table 4.7 presents the results obtained by the 16 participating systems – including OpAL- in the SemEval 2010 Task 18 competition. As it can be seen in this table, the system ranked fifth, with a Micro accuracy of 0.76037 and sixth, with a Macro accuracy of 0.7037. The data is reproduced from Wu and Jin (2010).

System	Micro Acc.(%)	Macro Acc.(%)
YSC-DSAA	94.20	92.93
HITSZ_CITYU_1	93.62	95.32
HITSZ_CITYU_2	93.32	95.79
Dsaa	88.07	86.20
OpAL	76.04	70.38
CityUHK4	72.47	69.80
CityUHK3	71.55	75.54
HITSZ_CITYU_3	66.58	62.94
QLK_DSAA_R	64.18	69.54

System	Micro Acc.(%)	Macro Acc.(%)
CityUHK2	62.63	60.85
CityUHK1	61.98	67.89
QLK_DSAA_NR	59.72	65.68
Twitter Sentiment	59.00	62.27
Twitter Sentiment_ext	56.77	61.09
Twitter Sentiment_zh	56.46	59.63
Biparty	51.08	51.26

Table 4.7: Results for the 16 system runs submitted (micro and macro accuracy)

Since the gold standard was not provided, we were not able to perform an exhaustive analysis of the errors. However, from a random inspection of the system results, we could see that a large number of errors was due to the translation – through which modifiers are placed far from the word they determine or the words are not translated with their best equivalent.

4.1.5. CREATING AN ANNOTATION SCHEME FOR FEATURE-BASED OPINION MINING

Subsequently to the initial research we have done in feature-based opinion mining, we proposed a new annotation scheme that is appropriated for the task and the elements that should be contemplated when tackling it (Balahur and Montoyo, 2009). Our motivations for designing and implementing this scheme, as well as creating new corpora that is labeled using this scheme are multiple.

The first one is the lack of a fine-grained opinion annotation scheme for reviews that takes into consideration the characteristics of this type of writing, both at the opinion versus fact level, as well as, within the opinion category, among the different methods to express it. Our first contribution is thus the definition of such an annotation scheme that encloses all the elements that should be considered at the time of opinion mining.

Further on, in order to evaluate systems implementing the paradigm of feature-based opinion mining, there are two available corpora, which are small, only concentrate on electronic products and are annotated at a sentence level, using a simple scheme under the form of *[feature name, polarity, value]*. Our second contribution is annotating a corpus of 100 reviews in English of different product categories (ranging from electro domestics to restaurants and books), using the created annotation scheme. We make this corpus available to further research.

Thirdly, we evaluate and validate our corpus annotations for fact versus opinion, using n-gram similarity to the annotated phrases and further classify opinionated sentences according to the polarity expressed, using n-gram similarity to the annotated sentences.

Finally, our contribution resides in proposing a method to use textual entailment in the opinion mining task. As far as we are aware of, such research has not been done so far. From the categories on which stars are given, we generate short positive and negative phrases (e.g. Ease of Use – “This is easy to use” versus “This is not easy to use.”) To these short hypotheses that we want to verify, we add examples of positive and negative phrases from the annotated corpus, which are related to the category. We test the entailment relation in a window of three consecutive phrases. The results obtained are encouraging.

OPINION ANNOTATION SCHEME

In order to train and test a system performing feature-based opinion mining, an annotated corpus is needed. However, in this field, of feature-based opinion mining, there are just two corpora available, annotated in a simple manner (for each sentence, the feature mentioned and the associated positive/negative score), both containing a small number of products and all pertaining to the same category (Hu and Liu, 2004; Ding et al., 2008). Another corpus developed for the more general opinion field is the Multi-Perspective Questioning Answering (MPQA) one (Wiebe et al., 2005), separating among the subjective and objective aspects of the annotated texts; however, this corpus only contains newspaper articles and does not take into consideration the aspects of product characteristic annotation that we are mostly interested in within the feature-based opinion mining task.

Starting with this observation, we decided to develop and apply an annotation scheme that would be able to capture all the aspects involved in the opinion mining process. Further on, we could use this annotated information for three experiments – fact versus opinion sentence classification, polarity classification for opinionated sentences and automatic feature detection.

As in the case of MPQA, our annotation scheme is designed for the integration within the GATE (General Architecture for Text Engineering)[17] framework. The annotation elements and their attributes are described using XML Schema files.

We describe four types of annotations: fact, opinion, feature expression and modifier.

The first element – ***fact*** (Figure 4.4)- was created for the labeling of tokens and phrases containing factual information. The attributes we defined for this element are ***source*** (writer, a quote etc.), ***target*** (name of product capability it refers to), ***id*** (an identifier given by the annotator for future reference), ***feature*** (name of the feature that the fact describes), ***type*** (direct, indirect or implicit), ***POS*** (part of

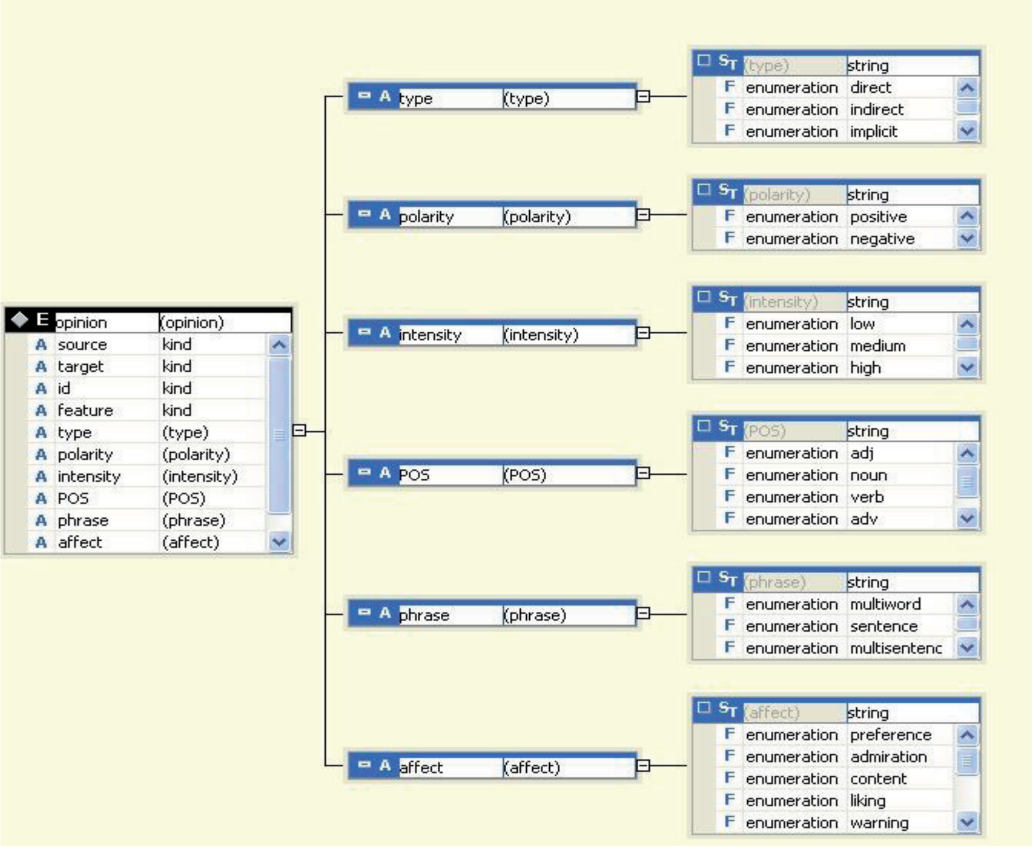
[17] http://gate.ac.uk

speech, for factual information expressed in individual tokens, which can be *adjective*, *noun*, *verb*, *adverb* or *preposition*), ***phrase*** (if the factual information is expressed in more than one token, a sentence or a group of sentences).

The second element – ***opinion*** – is presented in Figure 4.3. Apart from the attributes that are common to the fact element, this label contains the attributes ***polarity*** (positive or negative), ***intensity*** (degree of the opinion expressed, which can be *low*, *medium* or *high*) and ***affect*** (as results from the formulation of the opinion statement).

Another defined element – ***feature expression*,** with the same elements as the opinion tag, aims at identifying phrases that indirectly express a feature of the product under review (Figure 4.5).

Finally, the ***modifier*** element, whose structure is the same as for the opinion element, is included to spot the words whose use lead to a change in the polarity of the expressed opinion, but cannot be used alone to express an opinion (e.g. "*It's absurd")*. An example of annotation is given in Figure 4.6.



E opinion	(opinion)
A source	kind
A target	kind
A id	kind
A feature	kind
A type	(type)
A polarity	(polarity)
A intensity	(intensity)
A POS	(POS)
A phrase	(phrase)
A affect	(affect)

A type (type) — ST (type) string: F enumeration direct; F enumeration indirect; F enumeration implicit

A polarity (polarity) — ST (polarity) string: F enumeration positive; F enumeration negative

A intensity (intensity) — ST (intensity) string: F enumeration low; F enumeration medium; F enumeration high

A POS (POS) — ST (POS) string: F enumeration adj; F enumeration noun; F enumeration verb; F enumeration adv

A phrase (phrase) — ST (phrase) string: F enumeration multiword; F enumeration sentence; F enumeration multisentenc

A affect (affect) — ST (affect) string: F enumeration preference; F enumeration admiration; F enumeration content; F enumeration liking; F enumeration warning

Figure 4.3: Annotation scheme for the "opinion" element

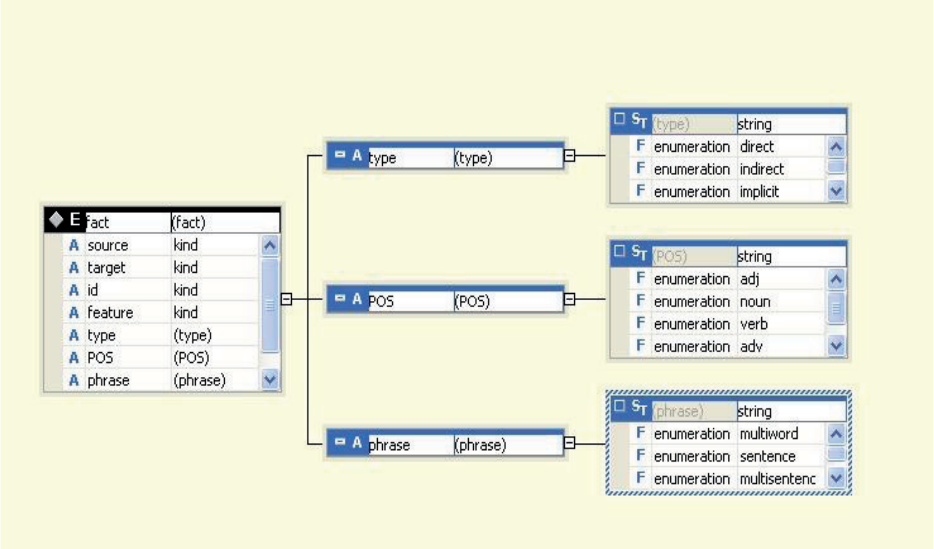


Figure 4.4: Annotation scheme for the "fact" element

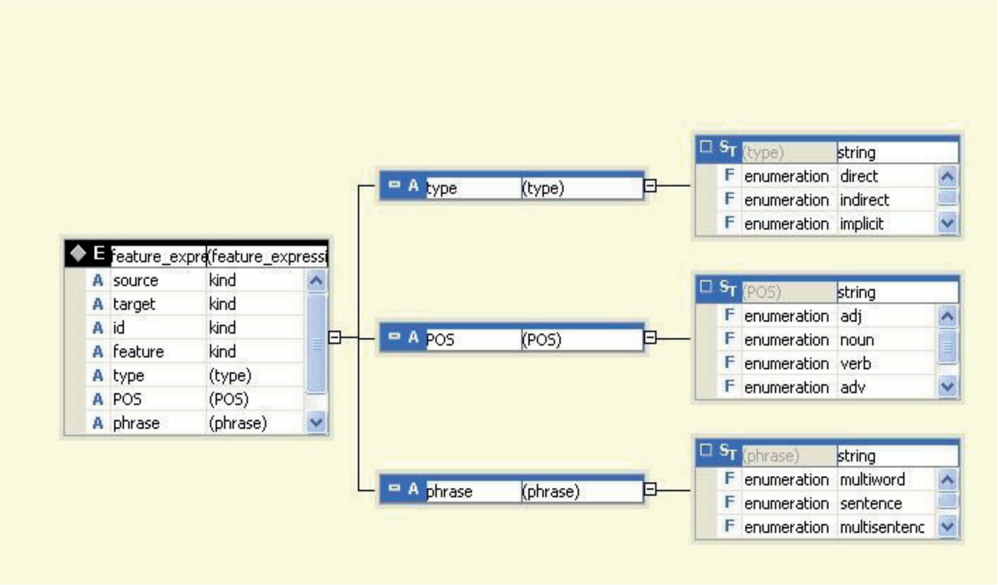
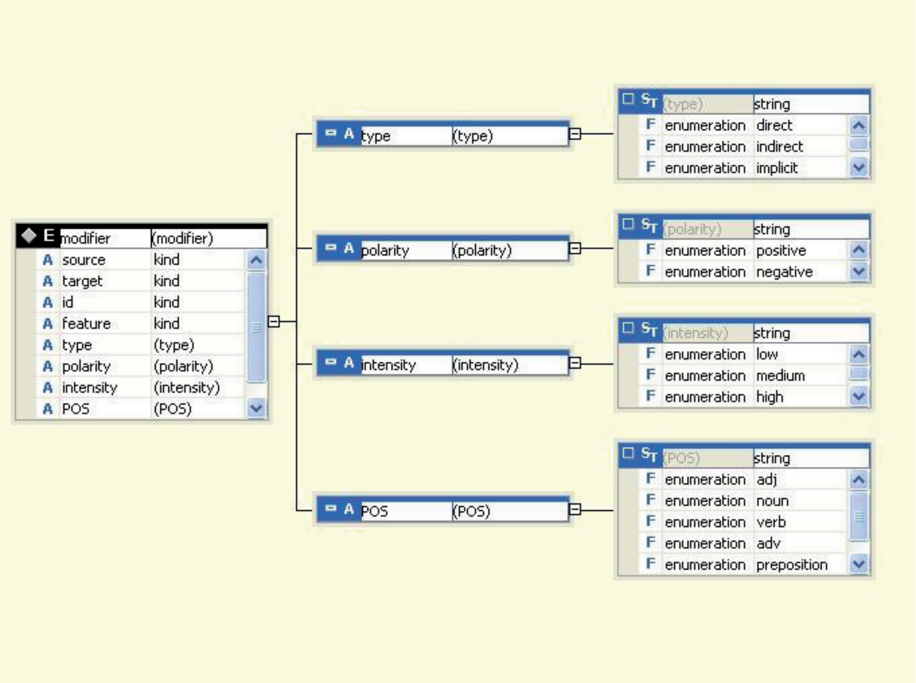


Figure 4.5: Annotation scheme for the "feature expression" element



E	modifier	(modifier)
A	source	kind
A	target	kind
A	id	kind
A	feature	kind
A	type	(type)
A	polarity	(polarity)
A	intensity	(intensity)
A	POS	(POS)

A	type	(type)

ST	(type)	string
F	enumeration	direct
F	enumeration	indirect
F	enumeration	implicit

A	polarity	(polarity)

ST	(polarity)	string
F	enumeration	positive
F	enumeration	negative

A	intensity	(intensity)

ST	(intensity)	string
F	enumeration	low
F	enumeration	medium
F	enumeration	high

A	POS	(POS)

ST	(POS)	string
F	enumeration	adj
F	enumeration	noun
F	enumeration	verb
F	enumeration	adv
F	enumeration	preposition

Figure 4.6: Annotation scheme for the "modifier" element

REVIEW CORPUS ANNOTATION

With the created scheme, we annotated a corpus containing 100 reviews, each containing approximately 1000 words (they range from 700 to 2000 words per review), on a high variety of "objects", pertaining to the categories of travel (resorts, hotels, restaurants and touristic attractions), home & garden (furniture, washing machines), electronics (laptops, PDAs, mobile phones, digital cameras), computers and software, music, movies, books and cars. All these reviews were taken from the *epinions.com* site. The advantage in using this source is that reviewers tend to profoundly analyze each of the aspects of the products and the reviews are long and complicated.

The idea behind the high variety of topics was, on the one hand, to detect similar manners to express opinions and, on the other hand, to test our opinion mining approaches at the coarse grained level, against the categories on which the users can punctuate the object using "stars" and at the fine-grained level, using our annotations. Moreover, choosing to label reviews on such a diversity of "objects" allowed us to test the applicability of the schema at the level of the feature-based opinion mining paradigm (i.e. identify, for any type of object, be it location, event, product etc., its features and the opinion words that are used to describe it). The annotation was done by two non-native speakers of English, one with a degree in Computer Science and the second a Linguistics student.

An example of annotation is presented in Figure 4.7.

```
<modifier gate:gateId="46" source ="w" target="restaurant"
id="83" feature="" type="direct" polarity="positive"
intensity="medium" POS="" phrase="multiword"
affect="admiration">I was pleasantly surprised</modifier>
to learn that the food is <modifier gate:gateId="47"
source ="w" target="excellent" id="84" feature=""
type="direct" polarity="positive" intensity="medium"
POS="adverb" phrase="word"
affect="admiration">still</modifier> <opinion
gate:gateId="48" source ="w" target="food" id="85"
feature="length" type="direct" polarity="positive"
intensity="high" POS="" phrase="word"
affect="admiration">excellent</opinion>, and the staff
very <opinion gate:gateId="49" source ="w" target="staff"
id="86" feature="length" type="direct" polarity="positive"
intensity="high" POS="adjective" phrase="word"
affect="admiration">professional</opinion> and <opinion
gate:gateId="50" source ="w" target="staff" id="87"
feature="length" type="direct" polarity="positive"
intensity="high" POS="adjective" phrase="word"
affect="admiration"> gracious </opinion>.
```

Figure 4.7. Sample from the annotated corpus

As a result of the annotation process, we found some interesting phenomena that are worth mentioning and that justify our annotation schema:

1. References along the argumentation line must be resolved (e.g. *I went to another restaurant from this chain and they treated us horribly. These people were very nice.*). This is the reason for which our scheme contains the elements "*target"* and *"id"*. In this way, all factual or opinion expressions referring to an entity can be traced along the text.
2. An idea is many times expressed along a line of sentences, combining facts and opinions to make a point (e.g. *This camera belongs to the new generation. Quality has highly improved over the last generations*.). We did not have a problem with this phenomenon, since our annotation scheme was designed to allow the labeling of token or expression-level elements, as well as sentence and multi-sentence level.
3. Within a phrase, a reviewer presents facts in such a manner that the reader is able to extract the corresponding opinion (e.g. *They let us wait*

for half an hour in the lobby and finally, when they gave us the key, realized the room had not been cleaned yet.). While the sentence is purely factual in nature, it contains the phrases *"let us wait for half an hour in the lobby"* , *"the room had not been cleaned yet"*, which we annotated as *feature expressions* of *implicit* type.

4. There is an extensive use of conditionals within reviews. (e.g. *If you dare, buy it! It was great for two weeks until it broke!; If you've got shaky hands, this is the camera for you and if you don't, there's no need to pay the extra $50…*) . We consider the sentence containing the conditional expression a *modifier*.
5. There are many rhetoric-related means of expressing opinion (e.g. *Are you for real? This is called a movie?)*. We annotate these elements as implicit *feature expressions*.

At the time of performing our experiments, these phenomena are important and must be taken into consideration, since 27% of the annotated opinionated phrases in our corpus are composed of more than one sentence. More generally, these findings draw our attention upon the context in which opinion mining is done. Most of the work so far concentrated on sentence or text level, so our findings draw the attention upon the fact that more intermediate levels should also be considered.

EXPERIMENTS AND EVALUATION

The first experiment we performed aimed at verifying the *quality and constancy of the annotation* as far as fact versus opinion phrases are concerned and, within the opinionated sentences, the performance obtained when classifying among positive and negative sentences. In the first phase, we lemmatize the annotated sentences using TreeTagger[18] and we represented each fact and opinion phrase as a vector of characteristics, measuring the n-gram similarity (with n ranging from 1 to 4) and overall similarity with each of the individual corpus annotated sentences, tokens and phrases. We perform a ten-fold cross validation using the SVM SMO. The results for fact versus opinion and positive versus negative classifications are presented in Table 4.8.

	Precision	**Recall**	**Kappa**
Fact	0.72	0.6	0.53
Opinion	0.68	0.79	0.53
Positive	0.799	0.53	0.65
Negative	0.72	0.769	0.65

Table 4.8. Evaluation of fact vs. opinion and positive vs. negative classification

[18] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

In the second phase, we consider for classification only the phrases containing two sentences. As in the first phase, we represent each fact and opinion phrase as a vector of characteristics, measuring the n-gram similarity (with n ranging from 1 to 4) and overall similarity with each of the individual corpus annotated sentences and with all other phrases containing two sentences, then perform a ten-fold cross validation using SVM SMO. We summarize the results obtained in Table 4.9.

	Precision	**Recall**	**Kappa**
Fact	0.88	0.76	0.43
Opinion	0.74	0.89	0.43
Positive	0.84	0.71	0.68
Negative	0.89	0.92	0.68

Table 4.9. Evaluation of fact versus opinion and positive versus negative 2-sentences phrase classification

In the third phase, we consider for classification only the phrases containing three sentences. The fact and opinion phrases are represented as in the first experiments and a ten-fold cross validation using SVM SMO is done. The results are shown in Table 4.10.

	Precision	**Recall**	**Kappa**
Fact	0.76	0.6	0.80
Opinion	0.78	0.94	0.80
Positive	0.85	0.76	0.68
Negative	0.92	0.96	0.68

Table 4.10. Evaluation of fact versus opinion and positive versus negative 3-sentences phrase classification

From the results obtained, we can notice that using longer phrases, we obtain an improved classification performance in both fact versus opinion classification, as well as positive versus negative classification. The only drop in classification performance is in the case of longer factual phrases. We explain this by the fact that in many of the cases, these types of phrases contain descriptions of opinionated sentences or represent combinations of factual and opinion sentences (e.g. *"They said it would be great. They gave their word that it would be the best investment ever made. It seems they were wrong"*). The results show that our annotation is constant and that labeled elements present similarity among them; this fact can be used to automate the annotation, as well as use the labeled corpus for the training of an opinion mining system or for its evaluation. Evaluation proved that the annotation schema and approach are general enough to be employed for labeling of reviews on any product.

Subsequently to annotating the review corpus, we study the manner in which feature-based opinion mining can be done at a coarse level (Balahur and Montoyo, 2009), i.e., not discovering all the features of a product and the corresponding opinion words that are used to describe them, but the opinion on the main aspects of the “object” on which opinion is mined, which are also evaluated by the users with “stars” (from 1 to 5 stars, 1 being the lowest and 5 being the highest). Comparing against the number of stars given to these main product functionalities can also be a useful method to evaluate opinion mining systems.

The approach taken in this section is motivated by the fact that many products on review sites have a “star” assigning system associated. In this manner, people are given the opportunity to value the product in question, besides using opinionated phrases, by employing, for a number of default defined features, from one to five stars. We consider this to be useful information at the time of review mining, both for the fact that we can thus overcome the problem of automatically discovering the distinct product features, as well as the problem of evaluating our approach without having a previously annotated corpus on the specific product.

Textual entailment recognition (Dagan et al., 2006) is the task of deciding, given two text snippets, one entitled Text (T) and the other called Hypothesis (H), if H can be inferred from T.

For this task, we will use the information given by the stars category. We consider the given criteria and generate simple sentences (hypotheses) for each of the positive and negative categories. The idea behind this proposed method is to test whether a textual entailment system would be able to capture and better resolve the semantic variability given by opinionated text.

Our textual entailment system (Iftene and Balahur-Dobrescu, 2007) is based on the tree edit distance algorithm. Each of the sentences is dependency parsed using Minipar[19] and passed through Lingpipe[20] in order to detect and classify the Named Entities it contains. Subsequently, syntactic, lexical and semantic similarities are computed, starting from the root of the dependency tree, between all the tokens and their corresponding dependency links in the hypothesis and the tokens in the text using the DIRT[21] collection, eXtended WordNet[22] and a set of rules for context modifiers.

[19] http://www.cs.ualberta.ca/~lindek/minipar.htm
[20] http://alias-i.com/lingpipe/
[21] http://demo.patrickpantel.com/Content/LexSem/paraphrase.htm
[22] http://xwn.hlt.utdallas.edu/

From the annotated corpus, for each of the considered products, we selected the reviews containing sentences describing opinions on the criteria which users are also allowed to assess using the stars system. The categories which are punctuated with two or less stars are considered as negative and those punctuated with four or five stars are considered as having been viewed positively.

We generated hypotheses under the form *“Category is good”* and *“Category is nice.”*, *“Category is not good”* and *“Category is not nice.”*,e.g. *“The price was good.”,”The price was not good.” , “The food was good.”, “The food was not good.” “The view was nice.”, “The view was not nice.”.* In case no entailment was found with such built sentences, we computed entailment with annotated sentences in the review corpus. The results obtained are shown in Table 4.11.

Name of product	Stars Category	Accuracy
Restaurant	Price	62%
	Service	58%
	Food	63%
	View	53%
	Atmosphere	58%
Digital camera	Ease of use	55%
	Durability	60%
	Battery life	80%
	Photo quality	65%
	Shutter lag	53%
Washing machine	Ease of Use	60%
	Durability	72%
	Ease of Cleaning	67%
	Style	65%

Table 4.11. Polarity classification accuracy against the number of stars per category

As we can see from the obtained results, textual entailment can be useful at the time of performing category based opinion mining. However, much remains to be done at the level of computing semantic similarity between opinionated texts. Such work may include the discovery of opinion paraphrases or opinion equivalence classes.

4.1.6. BUILDING A RECOMMENDER SYSTEM USING FEATURE-BASED OPINION MINING

Having seen the manner in which opinion mining can be done at the level of features, we envisage a straightforward method to recommend products based on their fine-grained characteristics, the assessments made on these features and the preferences a user can express. Thus, in order to recommend a product for purchase, we present a method to compute the similarity between a product (whose features are summarized using the feature-based opinion mining and summarization system) and what is seen as the “perfect” product in the category.

For each product category, we consider a list containing the general and product-specific features. The “perfect” product within that category can thus be represented as a vector whose indices correspond to the list of features and whose values are all 1, signifying that all features are 100% positive. At this point, it is interesting to note that the semantics of “positive” and “negative” for the product category are given by the feature attributes we extracted for each of the categories (positive for size thus includes “small”, “tiny”, “pocket-fit” etc.).

In order to find recommendable products, we use the customer review summarization system presented for each product model and its corresponding collection of reviews. In this manner, we build vectors corresponding to the product models, whose indices will be the same as the ones of the “perfect” product, and whose corresponding values will be the percentage in which the feature is classified as positive by the summarization system. Finally, we compute the similarity between the each of the obtained vectors and the vector corresponding to the “perfect” product using the cosine similarity measure. We recommend the top 5 matching products. In order to better understand the process of recommendation, we will consider an example and suppose the user would like to buy a 4-Megapixel camera. There are around 250 available models on the market and for each model one can read an average of 10 customer reviews. Instead of having to read 2500 reviews, employing the presented system, being given the 5 best products, the user will only have to browse through 50 reviews, in case (s)he is not confident in the system classification; when the user is confident, (s)he has to read none.

The list of features for a 4 Megapixel camera is: (price, warranty, size, design, appearance, weight, quality, lens, viewfinder, optical zoom, digital zoom, focus, image resolution, video resolution, memory, flash, battery, battery life, LCD size, LCD resolution, accessories).

The vector associated to the “perfect” 4-Megapixel camera will have as indices the features in the above list and all corresponding values 1: v_{perf}(price)=1; v_{perf}(warranty) =1 and so on, in the order given by the list of features. After applying the customer review summarization system on other 4-Megapixel cameras, we obtain among others the vectors v_1 and v_2, corresponding to Camera1 4MP and Camera2 4MP. In this case:

v_1=(0.7,0.5,0.6,0.2,0.3,0.6,0.5, 0.5,0.7,0.8,0.7,0.8,0.4,0.3,0.3,0.7,0.6,0.3,0.8,0.4,0.4)
v_2 = (0.8,1, 0.7,0.2,0.2,0.5, 0.4,0.4,0.8,0.8,0.8,0.8,0.7,0.7,0.3,0.8,0.6,0.7,0.5,0.3,0.6)

Calculating the cosine similarity between v_1 and v_{perf} and v_2 and v_{perf}, respectively, we obtain 0.945 and 0.937. Therefore, we conclude that Camera1 4MP is better than Camera2 4MP, because it is more similar to the “perfect” 4-Megapixel camera model.

4.2. OPINION MINING FROM NEWSPAPER ARTICLES

4.2.1. INTRODUCTION

Subsequently to the experiments in feature-based opinion mining, our aim was to apply sentiment analysis to newspaper articles. As mentioned before, the task can be formulated in the context of any textual type. Nevertheless, given the peculiarities of the genre and the final aim of the sentiment analysis task, the requirements of a system that automatically processes the intended kind of text to extract opinions are different. Anticipating the following sections, this observation was confirmed in the experiments we performed with news data, blogs, political debates and microtext.

In a first approach, we started researching on appropriate methods to classify sentiment expressed in news, with the aim of including an opinion mining component to the Europe Media Monitor[23] family of applications (Balahur et al., 2009f). Such data is very different from product reviews in that sentiment is usually expressed much less explicitly. Bias or sentiment can be expressed by mentioning some facts while omitting others, or it can be presented through subtle methods like sarcasm (e.g.”*Google is good for Google, but terrible for content providers”*). Another major difference between the news and product reviews is that the target of the sentiment is much less concrete. A camera has very well-defined and easily identifiable features like weight, flash light, battery life, etc., but what are the *“features”* of a named entity such as a specific person or of an organization like the European Commission (EC)? As opposed to the “features” of products, which are parts or characteristics that are easily linkable through technical details or frequent mentioning, the “features” of more general topics, such as persons, events or organizations can be considered as sub-topics (e.g. administration, policy areas,

[23] http://emm.newsbrief.eu/overview.html

development of poorer regions, consumer protection, environment issues). Newspaper articles are much more complex than reviews on products, as they contain in their majority part factual descriptions of events and their participants. It is thus tricky to detect whether any negative sentiment detected refers to a specific organization we are interested in, to the main news, or to any other entity or topic mentioned in the context. Furthermore, the context in itself may have a specific sentiment associated to it (i.e. the news can be good or bad, independent on the sentiment that is expressed on the entities or the topic it is related to). Therefore, a clear distinction must be done in order to separate the content in which the entity we are interested in was mentioned (e.g. a natural disaster), and the sentiment expressed towards this entity in the given context (e.g. if the entity was helpful and offered support, or, on the contrary, if it ignored the gravity of the situation and did nothing). Observing these characteristics of newspaper articles, we can state that different tasks can be performed in their context and each must be treated in a different manner, using specific methods and resources. For example, classifying the context as good or bad news is different from classifying the opinion expressed on a specific entity participating in the news or classifying the sentiment expressed on the general context. Additionally, all these tasks are distinct from the one involving the detection of author or source bias.

Given the complexity of the phenomena identified in the preliminary analysis, our first aim when trying to perform opinion mining from newspaper articles was to delimit the scope of the analysis and define the task we are going to tackle.

Our first aim was to detect sentiment in quotations (direct reported speech). The reason for this is that the text in quotes is usually more subjective than the other parts of news articles, where sentiment is either expressed less, or it is expressed less explicitly. We also know for quotes who the person is that made the statement (referred to as the source of the opinion statement) and – if the speaker makes reference to another entity within the quotation – we have a clue about the possible target (or object) of the sentiment statement. (e.g. *Steinmeier said: "I think we can conclude that there is a fresh wind in NATO, and also, hopefully, a new atmosphere of cooperation,"*). In the first experiments, we will thus propose different methods to classify quotations depending on whether or not they are subjective and to determine the polarity of the sentiment expressed in the identified subjective quotations. Unlike full articles, quotations are relatively short and often are about one subject. However, they contain a variety of interesting phenomena, such as the combination of a short, factual summary of the event or what the "target" did or a general view on the problem, as well as the opinion or position of the "source" on this fact description (e.g. *"It is a tough battle and those who perceive us as competitors are not going to roll over and play dead. But again, both Branson and Fernandes are battle scarred and, with a song and a prayer, and lots of hard work, I believe we shall prevail"*).

Another interesting aspect concerns the presence of various possible “targets” in the quote, on which antonymic opinions are expressed (e.g. *“How can they have a case against Fred, when he didn’t sign anything?”*). Moreover, they contain a larger scale of affective phenomena, which are not easily classifiable when using only the categories of positive and negative: warning (e.g. *“Delivering high quality education cannot be left to chance!”*), doubt (e.g. *“We don’t know what we should do at this point”)*, concern, confidence, justice etc. (where doubt is generally perceived as a negative sentiment and confidence as a positive one).

The aim we have is to determine the attitude polarity (tonality of the speech), independent of the type of news, interpreting only the content of the text and not the effect it has on the reader.

4.2.2. INITIAL EXPERIMENTS

For our first experiments, we chose a set of 99 quotes, on which agreement between a minimum of two annotators could be reached regarding their classification in the positive and negative categories, as well as their being neutral/controversial or improperly extracted. The result of the grouping was a total of 35 positive, 33 negative, 27 neutral/controversial and 4 improperly extracted quotes. We used this dataset to comparatively analyze the different possible methods and resources for opinion mining and we explored the possibility to combine them in order to increase the accuracy of the classification.

The first approach is based on a “bag of words” – the use of different lexicons containing positive and negative words. The second approach contemplates measuring similarity to annotations extracted from existing corpora and machine learning.

A. BAG-OF-WORDS APPROACH

At the present moment, there are different lexicons for affect detection and opinion mining. The aim in the following evaluation is to test the different resources in the quote classification scenario and assess the quality and consistency of these lexicons.

Each of the employed resources were mapped to four categories, which were given different scores – positive (1), negative (-1), high positive (4) and high negative (-4). The assignment of these values was based on the intuition that certain words carried a higher affective charge and their presence should be scored accordingly. Our intuition was supported by experiments in which we used just the positive and negative categories and that scored lower.

The polarity value of each of the quotes was computed as sum of the values of the words identified; a positive score leads to the classification of the quote as positive, whereas a final negative score leads to the system classifying the quote as negative. The resources used were: the JRC lists of opinion words, WordNet Affect (Strapparava and Valitutti, 2004), SentiWordNet (Esuli and Sebastiani, 2005), MicroWNOp (Cerini et al., 2007). WordNet Affect categories of anger and disgust were grouped under high negative, fear and sadness were considered negative, joy was taken as containing positive words and surprise as highly positive; SentiWordNet and MicroWNOp contained positive and negative scores between 0 and 1 and in their case, we mapped the positive scores lower than 0.5 to the positive category, the scores higher than 0.5 to the high positive set, the negative scores lower than 0.5 to the negative category and the ones higher than 0.5 to the high negative set.

As a filtering step, we first classified the quotes based on the presence of "subjectivity indicators", using the Opinion Finder lexicon (Wilson et al., 2005). The subjective versus objective filtering had an accuracy of 0.89, as 2 of the positive and 5 of the negative quotes were classified as neutral.

We evaluated the approaches both on the whole set of positive and negative quotes, as well as only the quotes that were classified as "subjective" by the subjectivity indicators. Subsequently, we grouped together resources that tended to over-classify quotes as positive or negative, in an attempt to balance among their classification. Finally, we grouped together all the words pertaining to the different classes of positive, negative, high positive and high negative words belonging to all the evaluated resources. The results are shown in Table 4.12 (-S/O and +S/O indicate absence and presence, respectively, of the subjectivity filtering):

Resource	-S/O	+S/O	P_{pos}	P_{neg}	R_{pos}	R_{neg}
JRCLists	X		0.77	0.3	0.54	0.55
		X	0.81	0.35	0.6	0.625
SentiWN	X		1	0	0.51	0
		X	1	0	0.54	0
WNAffect	X		0	1	0	0.51
		X	0	1	0	0.54
MicroWN	X		0.62	0.36	0.52	0.48
		X	0.73	0.35	0.57	0.53
SentiWN + WNAffect	X		0.22	0.66	0.42	0.45
		X	0.24	0.67	0.47	0.41
All	X		0.68	0.64	0.7	0.62
		X	0.73	0.71	0.75	0.69

Table 4.12: Results of the classification using the different opinion and affect lexicons

B. SIMILARITY APPROACH

In this approach we used two existing resources – the ISEAR corpus (Scherer and Walbott, 1997) - consisting of phrases where people describe a situation when they felt a certain emotion and EmotiBlog (Boldrini et al., 2009), a corpus of blog posts annotated at different levels of granularity (words, phrases, sentences etc.) according to the polarity of the sentiments and the emotion expressed.

In the first approach, we computed the individual quotes' similarity with the sentences belonging to each of the emotions in the ISEAR corpus, using Pedersen's Similarity Package[24], based on the Lesk similarity[25]. Subsequently, we classified each of the quotes based on the highest-scoring category of emotion. Table 4.13 presents the results:

Class	Joy	Fear	Anger	Shame	Disgust	Guilt	Sadness
Positive	8	7	1	3	3	5	8
Negative	6	7	1	5	8	2	4

Table 4.13: Results of the classification using the similarity scores with the ISEAR corpus

We consider as positive the examples which fell into the "joy" category and classify as negative the quotes which were labeled otherwise. The results are presented in Table 4.14:

P_{pos}	P_{neg}	R_{pos}	R_{neg}	Accuracy
0.22	0.82	0.58	0.5	0.514

Table 4.14: Results of the positive versus negative classification using the similarity score with the ISEAR corpus

EmotiBlog represents an annotation schema for opinion in blogs and the annotated corpus of blog posts that resulted when applying the schema. The results of the labeling were used to create a training model for an SVM classifier that will subsequently be used for the classification of opinion sentences. The features considered are the number of n-grams (n ranging from 1 to 4) and similarity scores with positive and negative annotated phrases, computed with Pedersen's Similarity Package. The approach was previously described by Balahur et al. (Balahur et al., 2009b). The evaluation results are presented in Table 4.15:

[24] http://www.d.umn.edu/~tpederse/text-similarity.html

[23]http://kobesearch.cpan.org/htdocs/WordNet-similarity/WordNet/ Similarity/lesk.htm

Class	Precision	Recall	F-measure
Positive	0.667	0.219	0.33
Negative	0.533	0.89	0.667

Table 4.15: Results of the classification using SVM on the EmotiBlog corpus model

From the results obtained, we can infer that the use of some of the resources leads to better performance when classifying positive or negative quotes (SentiWordNet versus WordNet Affect), and that the combined resources produce the best results when a vocabulary-based approach is used. Another conclusion is that previous subjectivity filtering indeed improves the results.

4.2.3. PRELIMINARY CONCLUSIONS FROM INITIAL EXPERIMENTS

From the results in Table 4.15, we can conclude that annotations about a specific topic cannot be applied to generic opinion mining on news. This confirms that open-domain opinion analysis is a more difficult problem than topic-specific sentiment classification and other sub-tasks defined in opinion mining, such as feature-based opinion mining. Experiments showed that simple bag-of-words approaches cannot reach a satisfactory level, even when large sets of words are employed. Most importantly, these preliminary experiments, even if they were performed on a very small dataset, have succeeded in shedding some light on the sentiment analysis – related phenomena in newspaper articles and the challenges that are associated to them.

Following these findings, we also realized that the task of sentiment analysis needs to be redefined in the context of news. This was done in a further effort, described by Balahur and Steinberger (2009). Experiments using the new definition provided have shown that indeed, when the task is clarified, both the annotation of newspaper article texts according to their sentiment has a better agreement, as well as the performance of the automatic processing increases.

4.2.4. REDEFINING SENTIMENT ANALYSIS FROM NEWSPAPER ARTICLES

Following this first set of experiments by Balahur and Steinberger (2009), we set the objective of annotating a larger corpus of quotations extracted from newspaper articles. We extracted a set of 1592 quotations in English which we set out to annotate with sentiment. The task was to decide whether a quotation was positive, negative or neutral.

After a first effort to annotate 400 quotations by 2 different persons, we realized that the inter-annotator agreement even for the short pieces of text was relatively low (below 50%). The following is an example of a quotation where annotators had difficulty to agree:

(1) Politician A said: *"We have declared a war on drugs"*.

While one annotator may feel that this is a positive statement as it shows action and efficiency to overcome an existing drug-related problem, another annotator may interpret this statement as being negative because (a) 'war' is negative and (b) the situation must be rather bad if an extreme reaction such as 'declaring war' is necessary. Depending on the background of the annotator, s/he may even want to argue in a certain context (c) that 'drugs' refers to soft drugs and that these are not a problem at all for society, in which case politician A's attitude would probably be considered as being misled and erroneous. While the source of the quotation (politician A) is clear, there is thus some confusion regarding what the target is. Is it the energetic and problem-solving (or erroneous) attitude of the politician (positive attitude), or is it the alleged fact that there is a drug problem (negative news). A further issue is the question whether the confidence expressed in the politician's statement should be considered as being positive, i.e. whether a statement such as "we will do something about this problem" should be considered a sentiment statement at all. It is clear that such a statement is intended to cause positive feelings towards the politician. Some existing sentiment or affect vocabulary lists do also include words like 'war' and 'mother' with the respective values negative or positive. By adding one level to this example:

(2) Person B mocked politician A's statement that *"We have declared a war on drugs"*.

Yet another interpretation is possible, namely that 'Person B' is the source and 'politician A' is the target, questioning also somehow the positive sentiment politician A wanted to create with his/her statement. The journalist writing the article may express his or her own opinion on this, as in:

(3) Person B unreasonably mocked politician A's statement that *"We have declared a war on drugs"*.

In this case, the journalist expresses negative sentiment towards 'Person B' for criticizing 'politician A'.

The chain could theoretically be continued: For instance, if the newspaper were a known defender of person A (and the corresponding) political party and attitudes, the whole statement (3) could be interpreted as sarcasm, inverting the negative sentiment of ‘Person B’ towards ‘Politician A’, and so on. While this is clearly a constructed example, our low inter-annotator agreement and the clarifying discussions showed that our initial sentiment annotation instructions were under-specified and left too much leeway for interpretation.

(4) Time called on the “War” on Drugs?

These are real examples of texts, where we can further notice difficulties. In this case, the journalist mocks the idea of delaying taking an action against drugs. Or the following example:

(5) Argentina and Mexico have taken significant steps towards decriminalising drugs amid a growing Latin American backlash against the US-sponsored “war on drugs”.

In this context, “US-sponsored” is the key expression towards understanding the negative opinion on the “war on drugs”.

For these reasons, we re-defined our task and subsequently annotated the whole set of 1592 quotations, after which the inter-annotator agreement was 0.81%.

PREVIOUS DEFINITIONS

In order to redefine the task, we first start by looking into the definitions that were given until this point.

Subjectivity analysis is defined by (Wiebe, 1994) as the “linguistic expression of somebody’s opinions, sentiments, emotions, evaluations, beliefs and speculations”. In her definition, the author was inspired by the work of the linguist Ann Banfield (Banfield, 1982), who defines as subjective the “sentences that take a character’s point of view (Uspensky, 1973)” and that present private states (Quirk, 1985) (that are not open to objective observation or verification) of an experiencer, holding an attitude, optionally towards an object. Subjectivity is opposed to objectivity, which is the expression of facts. As Kim and Hovy (2004) notice, opinion is subjective, but may not imply a sentiment. But what about our example of “war on drugs”? Can facts express opinions? Is there a difference between interpretations of facts at sentiment level and the direct expression of sentiments? Should we take them into consideration? Therefore, in our context, this definition did not help.

Esuli and Sebastiani (2006) define opinion mining as “a recent discipline at the crossroads of information retrieval and computational linguistics which is concerned not with the topic a document is about, but with the opinion it expresses”. This is a very broad definition, which targets opinions expressed at a document level. As we have shown before, news articles contain mentions of different persons and events, the topic in itself might involve a negative tonality and both the author of the text, as well as the facts presented or the interpretation they are given by the reader may lead to a different categorization of the document. So, this definition is not specific enough for us to understand what we should be looking for when annotating pieces of newspaper articles.

Dave et al. (2003) define an opinion mining system as one that is able to “process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good).” Opinion mining, in this context, aims therefore at extracting and analyzing judgments on various aspects of given products.

A similar paradigm is given by Hu and Liu (2004), which the authors entitle feature-based opinion mining. It is, however, not clear how statements such as “It broke in two days”, “The night photos are blurry”, that are actual fact information (according to the definition of subjectivity, they are verifiable) could be and should be annotated. Do they fall outside the goal of opinion mining? Since in our context, persons, organizations or events have no definable or inferable lists of features, this definition of the task does not work for us either.

Kim and Hovy (2005) define opinion as a quadruple (Topic, Holder, Claim, Sentiment), in which the Holder believes a Claim about the Topic, and in many cases associates a Sentiment, such as good or bad, with the belief. The authors distinguish among opinions with sentiment and opinions without sentiment and between directly and indirectly expressed opinions with sentiment. In this context, it does not remain clear how an example such as the “Local authorities have provided no help for the victims of the accident.” should be interpreted and why. Some might even argue that a statement they claim to be opinionated but with no sentiment – “Gap is likely to go bankrupt” (which would probably be interesting when assessing favorability in markets), has a sentiment and that sentiment is negative.

In the SemEval 2007 No. 14 Affective Text Task (Mihalcea and Strapparava, 2007), the systems were supposed to classify 1000 newpaper titles according to their valence and emotion contained. A title such as “Scientists proved that men’s perspiration raises women’s hormone levels” or “100 killed in bomb attack” were classified as negative. However, this is factual, verifiable information. Does this mean that when capturing the media sentiment, we should consider these results as being negative? Do these statements refer to a fact and are we interested in the information conveyed or in the sought effect? If so, which of these aspects would we include in a system doing sentiment analysis from newspaper articles?

In other approaches, capturing favorability versus unfavorability, support versus opposition, criticism versus appreciation, liking versus disliking, even bad versus good news classification were considered sentiment analysis.

However, at the moment of annotating sentiment in newspaper articles, we have seen that combining all these aspects together did not help to clear what the task was and how annotation should be done. Even in the case of quotes, which are short pieces of text where the source was known and the possible targets were identified, expressions of opinion that needed some kind of interpretation or knowledge of the situation fell short of agreement, due to personal convictions, background and so on.

REDEFINITION OF GUIDELINES FOR SENTIMENT ANNOTATION IN NEWSPAPER ARTICLES

We further on then present an annotation effort for newspaper quotes that shed light on the issue and helped define guidelines for labelling that led, from level of agreement of under 50%, to a level of agreement of 81%. We give some details on the gold-standard quotation collection we created according to these guidelines. Finally, we redefine the task of sentiment analysis in the news, capturing the different aspects of sentiment in text that we identified and pinpointing what exactly we expect a sentiment analysis system to discover from news under the different aspects.

REDEFINING THE TASK OF SENTIMENT ANNOTATION

Although some definitions of the task were proposed, none of them, as we have seen, could give an indication of the specific aspects that a system implementing opinion mining in a news context should contemplate. To clarify the task, we selected a collection of 1592 quotes (reported speech) from newspaper articles in English, whose source and target were known (their extraction patterns are designed in that scope) which we set out to annotate. Details on the length of the quotes are given in Figure 4.8.

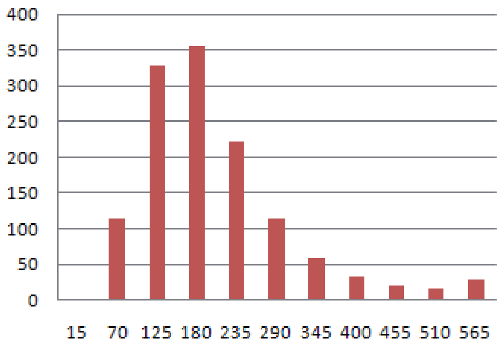


Figure 4.8: Histogram of the quotes' length

The first experiments had an agreement lower than 50%. Specifying that just the sentiment on the target should be annotated and separated from the good and bad news that was described led to an increase in the agreement up to 60%. We realized that by delimiting a few aspects, the task became much clearer. Following are the annotation guidelines we used:

We are trying to decide whether the entity (person or organization) in the text snippet is being talked about in a positive (POS) or in a negative light (NEG), or if the statement is rather objective/neutral (OBJ). We thus distinguish three cases of sentiment: two cases of subjectivity, in which case we can directly indicate the polarity (POS, NEG), and the case of non-subjectivity, objectivity or neutrality (OBJ). OBJ is the default, so no need to label neutral/objective examples.

Here are some more clarifications that may help:

1. *If you can, try not to make use of your world knowledge, such as the political views of the entity. If you cannot decide without knowing the political views of the entity, just leave it neutral/objective (OBJ).*
2. *It may help to imagine that you are the one being talked about: would you like or dislike the statement without using your world knowledge?*
3. *We are not interested in knowing whether the whole piece of text is positive or negative, but exclusively the sentiment expressed towards the entity.*
4. *Another hint to identify subjectivity – those news items whose content cannot be verified and whose content is expressly changed to induce a negative/positive opinion should be annotated as positive or negative. E.g. "Negotiations with Turkey have been delayed" – factual (objective from a sentiment point of view) vs. "EU stalls negotiations with Turkey" – (subjective, negative sentiment).*
5. *Note that, in the text snippet "X supported Y for criticizing Z", there is*

negative sentiment towards Z from X and Y, but positive sentiment between X and Y.

6. *Please try to separate good news vs. bad news from the sentiment expressed. We should NOT annotate the good versus bad content of the news. E.g. if the news talks about 7000 people being killed by a bomb, the news is factual/objective (OBJ), even if there is negative news content.*
7. *Expressions of attitude: "EU is willing to make efforts to prevent this from becoming a crisis" (This shows positive attitude, i.e. positive sentiment, POS); On the other hand, the sentence "EU sent help to the earthquake-affected Aquila citizens" is objective from a sentiment point of view (OBJ).*
8. *Should there be both positive and negative statements in the snippet, please consider the statement to be objective (OBJ). (strictly speaking, it would be subjective, but balanced; but we are not trying to distinguish this case).*
9. *It is certain that there will be many cases of doubt. In case of doubt, just leave the example un-annotated (neutral/objective, OBJ).*

The original data set we decided to annotate contained 1592 quotes extracted from news in April 2008. The average final agreement was 81%, between 3 pairs of two annotators each.

	Number of quotes	**Number of agreed quotes**	**Number of agreed negative quotes**	**Number of agreed positive quotes**	**Number of agreed objective quotes**
	1592	1292	234	193	865
Agreement		81%	78%	78%	83%

Table 4.16: Results of the annotations in terms of agreement per class of sentiment

The result of the annotation guidelines and labeling process what a corpus in which we agreed what sentiment was and was not in our case. The number of agreed sentiment-containing quotes was one third of the total number of agreed quotes, showing that only clear, expressly stated opinion, which required no subjective interpretation from the annotator's part was done.

The result of our labeling showed that in the case of newspapers, it is mandatory to distinguish between three different "components": the author, the reader and the text itself (Figure 4.9).

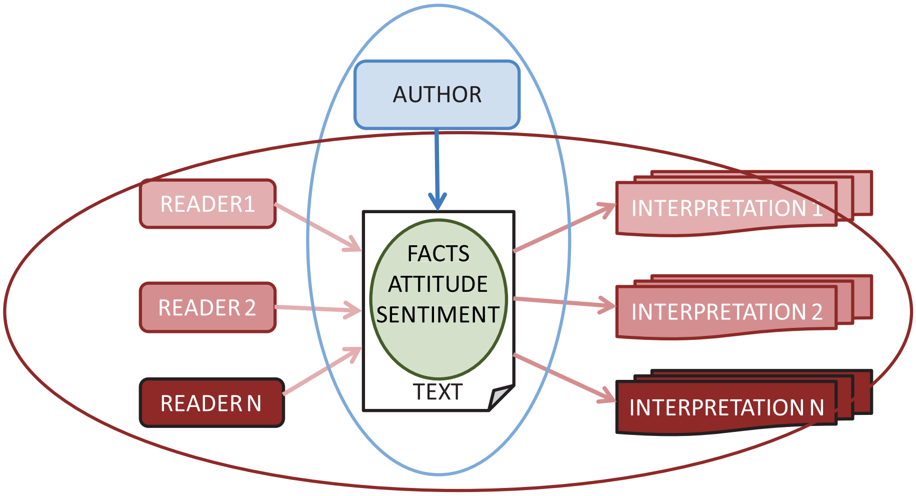


Figure 4.9: A 3-component view on sentiment expression – author, text, reader

While *the author* might convey certain opinions, by omitting or stressing upon some aspect of the text and by inserting their own opinion towards the facts, the spotting of such phenomena is outside the aim of sentiment analysis as we have defined it (and done by perspective determination, or news bias research). From the *reader's point of view*, the interpretations of the text can be multiple and they depend on the personal background knowledge, culture, social class, religion etc. as far as what is normal (expected) and what is not are concerned. Lastly, the opinion stated *strictly in the text* is the one that one should concentrate on at this level, being expressed directly or indirectly, by the target, towards the source, with all the information needed to draw this conclusion on polarity present in the text.

From the author and the reader's perspective and not from the text's pure informational point of view, opinion is conveyed through facts that are interpretable by the emotion they convey. However, emotions are not universal in their signification. They are determined socially, culturally and historically. There are general emotions, but most of the times they relate to the norms, their significance and the cultural environment. Emotions imply an evaluation, which is both cognitive and affective, of a behavior, with respect to a norm and the mutual expectation it raises. Some norms are common sensical and overall accepted and understood. Normative expectations link the behavior (reaction) to a meaning and on this ground, by the significance it is given. From the reader's point of view, sentiment analysis would be defined as *the assessment of a "target", based on its characteristics and factual information related to it, according to whether or not the results of the assessments are "according to" or "against" the "norm".*

From the author's point of view, news bias or perspective determination should be concerned with discovering the ways in which expression of facts, word choice, omissions, debate limitations, story framing, selection and use of sources of quotes and the quote boundaries, for example, conveys a certain sentiment or not. The sentiment content of the text, finally, is what is expressly stated, and not what is left to be understood between the lines. Although pragmatics, through the speech- act or other theories would argue there is no text that has no intended meaning, the sentiment or factual information conveyed is different from reader to reader and can thus not be done at a general level, as sentiment analysis intends to. For example, the text "The results of the match between Juventus Torino and Real Madrid last night are 3-0." would maybe be interpreted as something positive, a motive for pride in an Italian newspaper, it would be a negative, sad thing if reported by a Spanish source, it would be bad or good depending on whether or not an interested reader were pro or against the two teams and it would constitute just factual news from the strict point of view of the text. Given these three views one must be aware of at the time of constructing a sentiment analysis system for news, we can see that the task becomes much clearer and the agreement at the time of annotating texts, implementing and evaluating systems is higher.

Should one want to discover the possible interpretations of texts, sources' and readers' profiles must be defined and taken into consideration, for a whole understanding of the possible sentiment effects text has or is intended to have, and not just a general, often misunderstood one.

At this moment, having the tasks clearly defined, we have started experimenting with adequate methods to perform sentiment analysis considering these insights.

4.2.5. EXPERIMENTS APPLYING THE NEW TASK DEFINITION OF SENTIMENT ANALYSIS FROM NEWSPAPER ARTICLES

EXPERIMENTAL SETUP

In order to measure the impact of our defined task, we performed different experiments on the set of 1292 quotes on which agreement has been reached (Balahur et al, 2010d). Out of these 1292 quotations, the target was successfully identified by the sentiment analysis system in 1114 quotes (direct mentions of the *target* through the name or its title). The baseline we compare against is the percentage of quotes pertaining to the largest class of quotes – objective, which represents 61% of our corpus.

According to the approach we settled on, we wanted to make sure that: a) we estimate the opinion on the target of the quote (by computing the opinion in windows of words between the mentions of the entity), b) we eliminate the bad versus good news content (by eliminating those words which are both sentiment-bearing words and words that are part of EMM category definitions, from now on called *category words*). Given that we are faced with the task of classifying opinion in a general context, we employed a simple, yet efficient approach, presented in (Balahur et al., 2009f).

At the present moment, there are different lexicons for affect detection and opinion mining. In order to have a more extensive database of affect-related terms, in the following experiments we used *WordNet Affect* (Strapparava and Valitutti, 2004), *SentiWordNet* (Esuli and Sebastiani, 2006), *MicroWNOp* (Cerini et al, 2007). Additionally, we used an in-house built resource of opinion words with associated polarity, which we denote by *JRC Tonality*. Each of the employed resources was mapped to four categories, which were given different scores: positive (1), negative (-1), high positive (4) and high negative (-4). The score of each of the quotes was computed as sum of the values of the words identified around the mentions of the entity that was the target of the quote, either directly (using the name), or by its title (e.g. Gordon Brown can be referred to as "*Gordon"*, as "*Brown"* or as "*the British prime-minister*")[26]. The experiments conducted used different windows around the mentions of the target, by computing a score of the opinion words identified and eliminating the words that were at the same time opinion words and category words (e.g. *crisis*, *disaster*).

EVALUATION RESULTS

Table 4.17 presents an overview of the results obtained using different window sizes and eliminating or not the category words in terms of *accuracy* (number of quotes that the system correctly classified as positive, negative or neutral, divided by the total number of quotes).

As it can be seen, the different lexicons available performed dramatically different and the impact of eliminating the alert words was significant for some resources or none for others, i.e. in those cases where there were no category words that coincided with words in the respective lexicon.

[26] For the full details on how the names and corresponding titles are obtained, please see (Pouliquen and Steinberger, 2009).

Word window	W or W/O Alerts	JRC Tonality	MicroWN	WNAffect	SentiWN
Whole text	W Alerts	0.47	0.54	0.21	0.25
	W/O Alerts	0.44	0.53	0.2	0.2
3	W Alerts	0.51	0.53	0.24	0.25
	W/O Alerts	0.5	0.5	0.23	0.23
6	W Alerts	0.63	0.65	0.2	0.23
	W/O Alerts	0.58	0.6	0.18	0.15
6	W Alerts	0.82		0.2	0.23
	W/O Alerts	0.79		0.18	0.15
10	W Alerts	0.61	0.64	0.22	0.2
	W/O Alerts	0.56	0.64	0.15	0.11

Table 4.17: Accuracy obtained using different lexicons, window sizes and alerts

As we can see, computing sentiment around the mentions of the entity in smaller window sizes performs better than computing the overall sentiment of texts where the entities are mentioned. From our experiments, we could notice that some resources have a tendency to over-classify quotes as negative (WordNet Affect) and some have the tendency to over-classify quotes as positive (SentiWordNet). We have performed evaluations using combinations of these four lexicons. The best result we obtained were using the combination of JRC Tonality and MicroWN, on a window of 6 words; in this case, the accuracy we obtained was 82%. As we can see, the majority of the resources used did not pass the baseline (61%), which shows that large lexicons do not necessarily mean an increase in the performance of systems using them.

ERROR ANALYSIS

Subsequently to the evaluation, we have performed an analysis of the cases where the system fails to correctly classify the sentiment of the phrase or incorrectly classifies it as neutral.

The largest percentage of failures is represented by quotes which are erroneously classified as neutral, because no sentiment words are present to account for the opinion in an explicit manner (e.g. “We have given X enough time”, “He was the one behind all these atomic policies”, “These revelations provide, at the very least, evidence that X has been doing favours for friends”, “We have video evidence that activists of the X are giving out food products to voters”) or the use of idiomatic expressions to express sentiment (e.g. “They have stirred the hornet’s nest”).

Errors in misclassifying sentences as positive instead of negative or vice-versa were given by the use of irony (e.g. “X seemed to offer a lot of warm words, but very few plans to fight the recession”).

Finally, quotes were misclassified as positive or negative (when they should in fact be neutral) because of the presence of a different opinion target in the context (e.g. “I’ve had two excellent meetings with X”, “At the moment, Americans seem willing to support Y in his effort to win the war”, “everyone who wants Y to fail is an idiot, because it means we’re all in trouble”, “The chances of this strategy announced by X are far better than the purely military strategy of the past...”) or the use of anaphoric references to the real target.

All these problems require the implementation of specific methods to tackle them. Thus, firstly, the opinion lexicons should be extended to contain concepts which implicitly imply an assessment of the target because they are concepts we employ in our everyday lives (e.g. “hunger, food, approval”). Secondly, expressions that are frequently used in a language to describe “good” and “bad” situations have to be added to the opinion lexical (e.g. “stir the hornet’s nest”, “take the bull by the horns”). Irony is difficult to detect in text; however, when dealing with a larger context, the polarity of such pieces of text could be determined in relation to that of the surrounding sentences. Further on, we are researching on methods to determine the target of the opinion using Semantic Roles; thus, the judgement on the opinion expressed can be improved. Finally, resolving co-reference using a standard tool should in theory lead to a higher performance of the opinion mining system. However, in practice, from our preliminary experiments, the performance of the opinion mining system decreases when employing anaphora resolution tool.

These improvements are studied along the next chapters.

4.2. . CONCLUSIONS ON SENTIMENT ANALYSIS FROM NEWSPAPER ARTICLES

As we have seen, mining sentiment from newspaper articles is a different task from feature-based opinion mining and summarization, due to the complexities of the textual genre. We have seen that there is a need to clearly define, before the annotation is done, what the source and the target of the sentiment are, subsequently separate the good and bad *news content* from the good and bad *sentiment* expressed on the target and, finally, annotate only clearly marked opinion that is expressed explicitly, not needing interpretation or the use of world knowledge. We have furthermore seen that there are three different possible views on newspaper articles – author, reader and text – and they have to be addressed differently at the time of analysing sentiment. We have performed experiments in this direction, by using categories to separate good and bad news content from the opinionated parts of the text. We also evaluated our approach using different lexicons in diverse combinations, and word windows.

We have shown that this simple approach produces good results when the task is clearly defined. The data is available for public use at:

http://langtech.jrc.ec.europa.eu/Resources/2010_JRC_1590-Quotes-annotated-for-sentiment.zip

Subsequently to these annotation efforts, we have continued to work on the creation of corpora for sentiment analysis for other languages. We have created a corpus of quotations extracted from newspaper articles in German containing 2387 quotes, based on the same annotation criteria. This resource is also publicly available upon request to the authors.

4.3. OPINION MINING FROM POLITICAL DEBATES

4.3.1. INTRODUCTION

Until now, we have seen different methods to tackle opinion mining from reviews, on the one hand, and quotations extracted from newspaper articles, on the other hand.

The tasks involved, as well as the methodology used to detect and classify sentiment within these two textual genres are very different. Reviews are written by one author and usually have only one target – the product in question and its features. In newspaper articles, we have defined the task of sentiment analysis by separating three different components – the author, the text and the reader, and the text into the context of the news and the sentiment expressed. The fact that our experiments were done on already-extracted quotations, whose source and target were priori given made the task more straight-forward, in the sense that the mentions of the target entity could be traced within the text.

However, if we are to extend the frame of texts considered and try to detect the sentiment on a general topic or event, from a text where multiple targets and sources of sentiments are present, the methods we presented until now would not be directly applicable (e.g. the sentiments expressed on a law that is submitted for approval by the parliament, its benefits and drawbacks for the community it targets).

As a result of the participation in the Opinion Pilot track at the Text Analysis Conference (TAC) 2008, with the system presented by Balahur et al. (2008), which will be detailed in Chapter 5 of this thesis, we have decided to investigate on the resources and methods that are appropriate to treat opinion in any textual genre and independently of the nature of the source or target of the opinions (i.e. whether they are persons, events, products etc.). In the Opinion Pilot track, opinion questions were asked on 25 different targets (persons, events, products, topics related to persons etc.), whose answers were to be found in a set of blogs. As in this scenario, a general opinion mining system must deal with many topics, ranging from products to brands, companies, public figures, news topics, etc., which may not be directly stated in the text as such. Therefore, when pursuing the goal of classifying opinions, one must first of all have a base system that is able to detect negative and positive opinion. To this aim, we propose a general opinion mining system (Balahur et al., 2009e), which is able to track the sentiment expressed on different topics mentioned in political debates. In order to further generalize the nature of the text considered, we have chosen to test our initial methods for general opinion mining on a corpus of political debates. Taking into consideration the corpus we have at hand, we study the manner in which opinion can be classified along dialogues, depending on the intervening speakers. We evaluate two methods to aggregate opinion scores in order to make a unique classification of opinions provided by a given speaker. While this type of classification was previously done by Thomas et al. (2006), their approach was dependent on the previous training on the same kind of data; our approach is data-independent.

Last, but not least, we study the possibility to determine the source of the opinion expressed taking into consideration its polarity and the affect words used in expressing the arguments. Since the corpus is already annotated with the party the speaker belongs to, we perform this classification among the two parties represented – democrat and republican. While this type of classification was previously approached by Mullen and Malouf (2006), the authors took into consideration the general vocabulary used, and not the attitude towards the topic per se and vocabulary related to it.

4.3.2. BACKGROUND

Although in the State-of-the-art chapter we have specified the directions of research in sentiment analysis in general, we will briefly comment on work that is related to this particular effort. Related work includes document-level sentiment analysis and opinion classification in political texts.

Research in sentiment analysis at a document level, relevant research was done by Turney et al. (2002), who first select important sentences based on pre-specified part-of-speech patterns, then compute the semantic orientation of adjectives and subsequently sum up this orientation to determine the document polarity.

Other related research was done by Pang et al. (2002), who employ machine learning techniques to determine the overall sentiment in user reviews. Additionally, Dave et al. (2003) propose classifying opinion on products based on individual opinions on product parts. Gamon (2004) that studies the problems involved in machine learning approaches and the role of linguistic analysis for sentiment classification in customer feedback data. Matsumoto et al. (2005) research on the impact of dependency analysis in sentiment analysis. Finally, Ng et al. (2006) analyze the role of linguistic knowledge sources in opinion mining.

On the other hand, research in sentiment analysis from political texts included classifying texts as conservative, liberal or libertarian (Mullen and Malaouf, 2006), placing texts on an ideological scale (Laver et al., 2003; Martin and Vanberg, 2007). Other authors proposed methods to represent opposing viewpoints of two parties in conflict (Lin et al., 2006). The corpus used in our research was first put forward and employed by Thomas et al. (2006). In the research presented, the authors investigate the possibility to determine support and opposition to the proposed legislation from the floor debates. They use subjectivity indicators to train a SVM classifier on part of the data and then employ a minimum-cut graph algorithm to determine the orientation of the opinions. They perform individual evaluations, first classifying individual speech segments and secondly classifying opinion depending on the speakers (assuming that a speaker will maintain the same opinion throughout the debate on a topic).

Our approach differs in many aspects to the ones taken in previous work. First, we employ a general algorithm to classify the individual speech segments of the different persons participating in the debates on each of the topics into positive and negative. We base our classification on similarity measures between the speech segments and the words pertaining to the categories of affect, opinion and attitude. In the first phase, we perform the classification without taking into consideration the target of the opinion expressed in the speech segments and without assuming any opinion consistency with respect to the speakers. The second classification, performed at speaker level, is done independently of the data. While we show the manner in which machine learning can be employed on our method to improve the results of the classifications, we discuss the implications this brings to the generality of the system.

4.3.3. EXPERIMENTS

The corpus we use in our experiments is made up of congressional floor debates and was compiled by Thomas et al. (2006). The corpus is available for download and research[27]. It is split into three sets: the development set, the training set and the test set. The first one contains 702 documents (one document corresponds to one speech segment) pertaining to the discussion of 5 distinct debate topics, the training set contains 5660 documents organized on 38 discussion topics and the test set contains 1759 documents belonging to 10 debates. The corpus contains three versions of these three sets, with the difference consisting in the removal of certain clues relating to the topic and the speaker referred to in the speech. The speech-

[27] http://www.cs.cornell.edu/home/llee/data/convote.html

segment file-naming convention, ###_@@@@@@_%%%%$$$_PMV is decoded as follows[28]:

1. *### is an index identifying the bill under discussion in the speech segment (hence, this number also identifies the 'debate' to which the speech segment belongs)*
2. *@@@@@@ is an index identifying the speaker*
3. *%%%% is the index for the page of the Congressional record on which the speech segment appears, i.e., a number from 0001 to 3268 corresponding to one of the original HTML pages that we downloaded from govtrack.us .*
4. *$$$ is an index indicating the position of the speech segment within its page of the Congressional record. Hence, for example, a file named 055_400144_1031004_DON.txt would be the 4th speech on the 1031st HTML page of the record.*
5. *'P' is replaced by a party indicator, D or R (or X if no corresponding party could be found). As mentioned in the paper, we purposely *did not* use this information in our experiments.*
6. *'M' is replaced by an indicator of whether the bill under discussion is mentioned directly in the speech segment, or whether it is only referenced by another speech segment on the same page. If the bill is directly mentioned in the current speech, the letter M appears in the file name; otherwise, the letter O appears.*
7. *'V' is replaced by a vote indicator, Y or N, which serves as the ground-truth label for the speech.*

In the following experiments, we only make use of the decoding at positions 1, 2, 5 and 7. We also need to mention, that out of the three variants in which the data is available, we chose to use the first stage of the data. Therefore, we can use the references that are annotated within the individual speech segments, which is useful for the third approach of the first task.

POLARITY CLASSIFICATION

The first experiment we performed was classifying the data on the basis of polarity of opinion. The first approach to this task was determining the polarity of the debate segments, taken individually. At this stage, we did not consider the information regarding the speaker. In order to perform this, we used the similarity measure

[28] Taken from the README file of the corpus.

given by Ted Pedersen’s Statistics Package[29] with affect, opinion and attitude lexicon.

The affect lexicon consisted of three different sources: WordNet Affect - (with 6 categories of emotion – joy, surprise, anger, fear, sadness, disgust), the ISEAR corpus (Scherer and Walbott, 1997) – that contains the 7 categories of emotion – anger, disgust, fear, guilt, joy, sadness and shame, from which stopwords are eliminated) and the emotion triggers database (Balahur and Montoyo, 2008; Balahur and Montoyo, 2008c; Balahur and Montoyo, 2008e)- which contains terms related to human needs and motivations annotated with the 6 emotion categories of WordNet Affect.

The opinion lexicon contained words expressing positive and negative values (such as “good”, “bad”, “great”, “impressive” etc.) obtained from the opinion mining corpus in (Balahur and Montoyo, and to which their corresponding nouns, verbs and adverbs were added using Roget’s Thesaurus.

Finally, the attitude corpus contains the categories of ”accept”, ”approval”, ”confidence”, ”importance”, ”competence”, “correctness”, “justice”, “power”, “support”, “truth” and “trust”, with their corresponding antonymic categories – “criticism”, ”opposition”, “uncertainty”, “doubt”, ”unimportance”, “incompetence”, ”injustice”, “objection”, ”refusal” , ”incorrectness”.

After obtaining the similarity scores, we summed up the scores pertaining to positive categories of emotion, opinion and attitude and the negative categories, respectively. Therefore, the general positive score was computed as sum of the individual similarity scores for the categories of “joy” and “surprise” from the affect category, the “positive values” of the opinion lexicon and the “accept”, “approval”, “competence”, “confidence”, “correctness”, “justice”, “power”, “support”, “trust” and “truth”. On the other hand, the general negative score was computed as sum of the “anger”, “fear”, “sadness”, “shame” from the affect categories, the “negative values” of the opinion lexicon and the “criticism”, “opposition”, “uncertainty”, “doubt”, “unimportance”, “incompetence”, “injustice”, “objection”, “refusal” and “incorrectness” categories of the attitude lexicon. The first classification between negative and positive speaker segments was done comparing these two resulting scores and selecting the higher of the two as final value for polarity. We evaluated the approach on the development, training and test sets (Classification 1).

On the other hand, we employed the scores obtained for each of the emotion, opinion and attitude categories, as well as the combined scores used for classifying in the first step for the training of an SVM classifier, using the development and training sets. We then tested the approach on the test set solely. Due to the fact that in the affect category there are more negative emotions (4) than positive ones (only

[29] http://www.d.umn.edu/~tpederse/text-similarity.html

2), we chose for classification only the two strongest emotions according to the similarity scores found in the first approach. Those two categories were “fear” and “anger”. The results are presented under Classification 2.

Further, we parsed the speaker segments using Minipar[30], in order to determine possible dependency paths between words pertaining to our affect, opinion or attitude lexicon and the topic under discussion or mentioning of another speaker. Our guess was that many of the speech segments that had been classified as negative, although the ground-truth annotation had them assigned a positive value, contained a negative opinion, but not on the topic under discussion, but on the opinion that was expressed by one of the anterior speakers. Therefore, the goal of our approach was to see whether the false negatives were due to the classification method or due to the fact that the object on which the opinion was given was not the one we had in mind when classifying. In order to verify our hypothesis, we extracted from the files in which the opinion words from the files with similarity higher than 0 appeared and sought dependency relations between those words and the mention of a speaker (based on the number assigned) or the words describing the topic discussed – marked in files in which this names appear, the words “bill”, ”legislation”, “amendment” and “measure”. Affect, opinion or attitude words to which no relation was found to the mentioned topic or a speaker were discarded. In this approach, we did not use anaphora resolution, although, theoretically, it could help improve the results obtained. It would be interesting to study the effect of applying anaphora resolution on this task.

The results of the classification are summed up under Classification 3. Figure 4.10 presents an example of the dependency analysis for one of the sentences in which an attitude word was identified. It can be see that the word “support” – pertaining to the attitude category, has a dependency path towards the name of the bill under discussion – “h.r. 3283”. Figure 4.11 shows a schematic overview of the first approach, with the resources, tools and methods employed therein.

```
> (
E0     (()    fin C *      )
1      (i     ~ N   2      s      (gov rise))
2      (rise ~ V   E0     i      (gov fin))
E2     (()    I N   2      subj  (gov rise)  (antecedent 1))
3      (in    ~ Prep       2      mod    (gov rise))
4      (strong      ~ A   5      mod    (gov support))
5      (support     ~ N   3      pcomp-n      (gov in))
6      (of    ~ Prep       5      mod    (gov support))
7      (h.r   ~ N   6      pcomp-n      (gov of))
8      (3283 ~ N   7      num    (gov h.r))
```

[30] http://www.cs.ualberta.ca/~lindek/minipar.htm

```
9      (,     ~ U   7     punc  (gov h.r))
10     (the   ~ Det 16    det   (gov act))
11     (united     ~ U   12    lex-mod    (gov United States))
12     (states     United States N   16    nn    (gov act))
13     (trade      ~ A   16    mod   (gov act))
14     (rights     right N     16    nn    (gov act))
15     (enforcement      ~ N   16    nn    (gov act))
16     (act   ~ N   7     appo  (gov h.r))
17     (.     ~ U   *     punc)
)
```

Figure 4.10: Minipar output for a sentence in topic 421 on act "h.r. 3283"

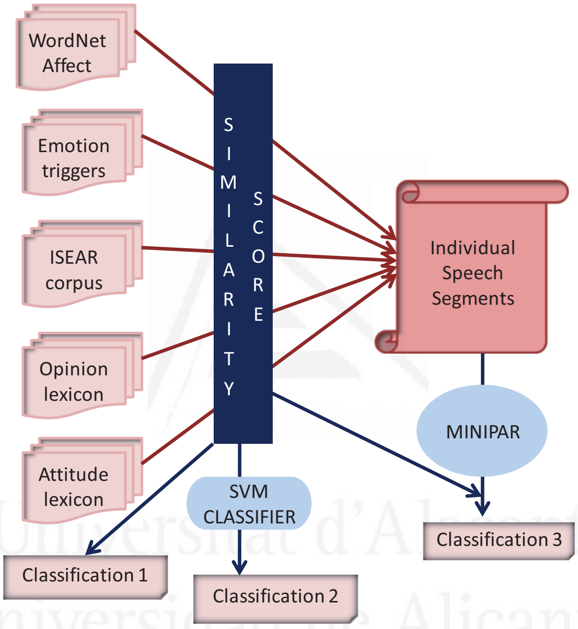


Figure 4.11: Resources and tools scheme for the first approach

The second approach on the data was aggregating the individual speaker segments on the same debate topic into single documents we denote as "speaker interventions". We then performed, on the one hand, a classification of these interventions using the sum-up of the scores obtained in the individual speech segments and, on the other hand, based on the highest score in each of the categories. Thirdly, we employed SVM to classify the speaker interventions using the aggregated scores from the individual text segments and the highest scores of the individual speaker segments, respectively. The training was performed on the development and training sets and the classifications (Classification 4 and Classification 5, respectively) were evaluated on the test set.

SOURCE CLASSIFICATION

The second experiment we performed was classifying the source of opinions expressed. In the following experiments, we used the fact that the corpus contained the name of the party the speaker belonged to coded in the filenames. The goal was to see whether or not we are able to determine the party a speaker belongs to, by taking into consideration the words used to express opinion on a given subject, the arguments (the words used within the argumentation) and the attitude on the subject in question.

Our hypothesis was that, for example, parties in favor of a certain piece of legislation will use both a set of words that are positively speaking on the matter, as well as a set of arguments related to the topic that are highlighting the positive side.

In order to perform this task, we used a clustering on the words pertaining to the affect lexicon, opinion lexicon and attitude lexicon, as well as the most frequent words appearing in the individual speaker segments of persons belonging to each of the two parties – Democrat and Republican.

As mentioned by Mullen and Malouf (2006), there are two problems that arise when intending to classify pertainance to a political party in a topic debate. The first one is the fact that when talking on a certain topic, all or most persons participating in the debate will use the same vocabulary. The second issue is that a certain attitude on a topic cannot reliably predict the attitude on another topic. Related to the first problem, we verify whether or not attitude towards a topic can be discriminated on the basis of the arguments given in support or against that topic, together with the affect, opinion and attitude lexicon used in connection to the arguments. As far as the second issue is concerned, we do not aim to classify depending on topic, but rather predict, on the basis of the arguments and affective words used, the party the speaker belongs to.

4.3.4. EVALUATION

We evaluated the approaches described in terms of precision and recall.

In order to exemplify the manner in which we calculated these scores, we present confusion matrices for all individual speech segments pertaining to the 5 topics in the development set.

The "yes" category includes the individual speech segments whose ground truth was "yes". The "no" category includes the individual speech segments that whose ground truth was "no".

The “positive” category includes the individual speech segments that the system classified as positive. The “negative” category includes the individual speech segments the system classified as negative.

	yes	**no**
positive	30	7
negative	22	16

Table 4.18: Confusion matrix for topic 199 from the development set

	yes	**no**
positive	45	2
negative	14	26

Table 4.19: Confusion matrix for topic 553 from the development set

	yes	**no**
positive	28	3
negative	15	29

Table 4.20: Confusion matrix for topic 421 from the development set

	yes	**no**
positive	44	3
negative	26	58

Table 4.21: Confusion matrix for topic 493 from the development set

	yes	**no**
positive	35	2
negative	24	26

Table 4.22: Confusion matrix for topic 052 from the development set

The following table shows the confusion matrix for the source classification, trained on the development set and tested using a sample of 100 documents from the test set, equally distributed among the Democrat and Republican Party.

	D	**R**
Classified D	29	21
Classified R	10	40

Table 4.23: Results for source classification 100 documents

We computed precision over positive classification as the number of individual text segments that our system classified as positive and that had the ground truth "yes" divided by the number of individual text segments that our system classified as positive and that had the ground truth "yes" summed with the number of individual text segments that our system classified as positive and that had the ground truth "no".

We computed precision over negative classification as the number of individual text segments that our system classified as negative and that had the ground truth "no" divided by the number of individual text segments that our system classified as negative and that had the ground truth "no" summed with the number of individual text segments that our system classified as negative and that had the ground truth "yes".

P_{pos}(199) = 30/37 = 0.81; P_{pos}(553) = 0.95; P_{pos}(421) = 0.90; P_{pos}(493)=0.93; P_{pos}(052)=0.94

P_{neg}(199) = 16/38 = 0.43; P_{neg}(553) = 0.65; P_{neg}(421) = 0.66; P_{neg}(493) = 0.78; P_{neg}(052)=0.52

We computed recall over positive classification as the number of individual text segments that our system classified as positive and that had the ground truth "yes" divided by the number of individual text segments that our system classified as positive and that had the ground truth "yes" summed with the number of individual text segments that our system classified as negative and that had the ground truth "yes".

We computed recall over negative classification as the number of individual text segments that our system classified as negative and that had the ground truth "no" divided by the number of individual text segments that our system classified as negative and that had the ground truth "no" summed with the number of individual text segments that our system classified as positive and that had the ground truth "no".

R_{pos}(199) = 30/52 = 0.57; R_{pos}(553) = 0.76; R_{pos}(421) = 0.65; R_{pos}(493) = 0.62; R_{pos}(052) = 0.59;

R_{neg}(199) = 16/23 = 0.70; R_{neg}(553) = 0.92; R_{neg}(421) = 0.90; R_{neg}(493) = 0.95; R_{neg}(052)= 0.92

We compute the accuracy score as the sum of the number of correct positive classifications and the number of correct negative classifications, divided by the total number of documents on a topic.

A(199) = 46/75 = 0.62; A(553) = 0.81; A(421) = 0.76;

A(493) = 0.77; A (052) = 0.70

The overall precision over positive classification is computed as average of all the precision scores of all the positive classifications. The precision over negative classifications is computed in the same manner.

The overall recall over positive classification is computed as average of all the recall scores of all the positive classifications. The recall scores over negative classifications is computed in the same manner.

The overall accuracy is computed as average of all the accuracy scores of all the topics.

	P_{pos}	P_{neg}	R_{pos}	R_{neg}	**Accuracy**
Development set	0.71	0.6	0.63	0.87	0.73
Training set	0.69	0.61	0.63	0.86	0.72
Test set	0.70	0.6	0.62	0.87	0.73

Table 4.24: Classification 1 (individual speaker segments) based on sums of similarity scores to affect, opinion and attitude lexicon categories

	P_{pos}	P_{neg}	R_{pos}	R_{neg}	**Accuracy**
Development set	0.72	0.69	0.69	0.85	0.77
Training set	0.71	0.68	0.68	0.85	0.76
Test set	0.70	0.67	0.68	0.84	0.76

Table 4.25: Classification 2 (individual speaker segments) using dependency parsing

	P_{pos}	P_{neg}	R_{pos}	R_{neg}	**Accuracy**
Test set	0. 75	0.63	0.66	0.88	0.78

Table 4.26: Classification 3 (individual speaker segments) based on SVM

	Accuracy
Development set	0.68
Training set	0.66
Test set	0.67

Table 4.27: Classification 4 (speaker interventions) based on sum-up of scores

	Accuracy
Development set	0.56
Training set	0.55
Test set	0.58

Table 4.28: Classification 5 (individual speaker segments) based on highest score

4.3.5. DISCUSSION AND CONCLUSIONS ON SENTIMENT ANALYSIS FROM POLITICAL DEBATES

As can be noticed from the scores obtained, the system has a rather balanced behavior on all topics in question. This is a positive fact, because the composition of the documents in the corpus was different, with as much as double the number of positive speaker interventions than negative ones.

We notice the tendency of the system to overly classify interventions as negative, a fact which is rectified by SVM learning, where better results are obtained. Dependency parsing was also found to help on the classification, improving the system's performance noticeably. SVM performed better than the initial method used for classification, but overall, the performance was lower than that obtained when performing dependency analysis.

We can also notice that when uniting speaker segments, the classification is done with better results. We believe this to be due to the fact that in the context of the larger text segments, more of the emotion categories have assigned a value above 0, and therefore more scores are important to the final result. Another factor to take into consideration when we analyze the results is the fact that speaker interventions were not equal as far as text length. From the post evaluation analysis, we found that the system performs better in classifying longer texts, to which more categories of emotion have similarity scores above 0.

It is interesting to note the fact that the categories that had the most importance at the time of classifying were those of "fear" and "joy" from the affect list – but given not by the lexicon in WordNet Affect, but from the categories found in the emotion triggers. As far as opinion source is concerned, the results shown in Table 4.23 demonstrate that a classification with 0.69% accuracy can be easily achieved using the affect and argument specific lexicon.

In order to tackle the issue of opinion mining from a more general perspective, we have proposed different methods to classify opinion from texts using affect, opinion and attitude lexica. We applied the proposed approaches to Congressional debates. We presented three methods to classify individual speech segments, the first based on a simple sum of similarity scores to the three lexicons used. We showed that applying dependency parsing and discovering the target of the opinion improved the initial classification, not only in the sense of the scores obtained, but also helped to balance between the results obtained for the positive and negative categories, respectively. Further, we showed that using SVM machine learning, we can improve the classification of the opinions, but that SVM performs worse than dependency analysis in the sense that it does not help improve the misclassification of positive opinions as negative ones. We showed that classification is dependent on the text length and that speaker interventions are better classified than individual speech segments alone, based only on the aggregated score obtained for each of the individual segment files. As far as opinion source is concerned, we showed that using affect, opinion and attitude lexicon in relation to the arguments given within the speech segments can help classify the source of opinion with 69% accuracy.

4.4. OPINION MINING FROM BLOGS

4.4.1. INTRODUCTION

Following the initial efforts to propose a general method for opinion mining and the results obtained, we realized that such an approach required not only appropriate methods, but new resources that are appropriately annotated, so that NLP systems could be trained on them.

As we have seen until now, opinions can be present at the level of individual words, in sentences, phrases and even in consecutive sentences. Moreover, while in texts such as reviews, the opinion is expressed most of the time by a single source and about one target, in other types of texts, such as newspaper articles, there are different components of the text that have to be analyzed separately. From the experiments we have done until this point and the analysis performed on the results, we can deduce that the requirements for a resource (annotated corpus) that is able to capture the phenomena involved in opinion expression so that an automatic system is able to detect it are the following ones. The resource should:

- Capture the difference between author intention, user interpretations and directly expressed opinion;
- Distinguish between opinion with sentiment and opinion without sentiment;
- Contain annotations of emotion, as a basic component of sentiment (so that experiments can be done to determine sentiment, even in the cases where it is present in objective sentences);
- Distinguish between opinion expressed in subjective sentences and opinion expressed through the mention of factual data, directly, indirectly or implicitly;
- Contain the annotation for opinion source and target (to avoid mixing the polarity of the context with the sentiment expressed on a specific entity);
- Track mentions of the source and target, whether they are present in text by anaphoric references, or topic-related mentions;
- Annotate opinion at the appropriate level – whether it is expressed through a word, a phrase, a sentence or various sentences;
- Capture the modalities of opinion expression both in traditional genres of texts, as well as in genres that are specific to the Social Web – e.g. blogs, forums, reviews, microblogs, social network comments; this includes the

possibility to annotate the style of writing, the language employed and the structure of writing and linking (e.g. misspellings, ungrammaticality, shortening of words and/or repetition of letters and punctuation signs, the use of colloquial expressions, “urban” acronyms, “smileys”).

The annotation scheme that allows for all these elements to be labeled is called EmotiBlog and it was proposed by Boldrini et al. (2009). The corpus that was annotated with this scheme contains blog posts in three languages: Spanish, Italian, and English about three subjects of interest, which in part overlap with the topics annotated in the MPQA corpus. The aim in choosing similar topics is that we can subsequently compare the results of the systems depending on the type of texts considered.

The first one contains blog posts commenting upon the signing of the Kyoto Protocol against global warming, the second collection consists of blog entries about the Mugabe government in Zimbabwe, and finally we selected a series of blog posts discussing the issues related to the 2008 USA presidential elections. For each of the abovementioned topics, we have manually selected 100 blog posts, summing up a total of 30.000 words approximately for each language.

4.4.2. THE EMOTIBLOG ANNOTATION MODEL

The *EmotiBlog* (Boldrini et al., 2009) annotation is divided into different levels, detailed in Table 4.29.

Element	Description
Objective speech	Confidence, comment, source, target.
Subjective speech	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Adjectives	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Adverbs	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Verbs	Confidence, comment, level, emotion, phenomenon, polarity, mode, source and target.
Anaphora	Confidence, comment, type, source and target.
Capital letter	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Punctuation	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Names	Confidence, comment, level, emotion, phenomenon,

Element	Description
	modifier/not, polarity, and source.
Reader Interpretation	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Author Interpretation	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Emotions	Confidence, comment; accept, anger, anticipation, anxiety, appreciation, bad, bewilderment, comfort, compassion, confidence, consternation, correct, criticism, disappointment discomfort, disgust, despondency, depression, envy, enmity, excuse, force, fear, grief, guilt, greed, hatred, hope, irony, interesting, important.

Table 4.29: EmotiBlog structure

For each element we are labelling the annotator has to insert his level of confidence. In this way, each label is assigned a weight that will be computed for future evaluations. Moreover, the annotator has to insert the polarity, which can be positive or negative, the level (high, medium, and low) and also the emotion this element is expressing. The phenomenon level describes whether the element is a saying or a colloquialism or a multi-word phrase.

As suggested by Balahur and Steinberger (2009), even if the writer uses an apparently objective formulation, he/she intends to transmit an emotion and a sentiment. For this reason we added two elements: reader and author interpretation. The first one is the impression/feeling/reaction the reader has reading the intervention and what s/he can deduce from the piece of text and the author interpretation is what we can understand from the author (politic orientation, preferences). Another innovative element we inserted in the model is the co-reference but just at a cross-post level. It is necessary because blogs are composed by posts linked between them and thus cross-document co-reference can help the reader to follow the conversations. We also label the unusual usage of capital letters and repeated punctuation. In fact, it is very common in blogs to find words written in capital letter or with no conventional usage of punctuation; these features usually mean shouts or a particular mood of the writer. Using *EmotiBlog*, we annotate the single elements, but we also mark sayings or collocations, representative of each language. Finally we insert for each element the source and topic.

4.4.3. EXPERIMENTS AND EVALUATIONS OF EMOTIBLOG ON ENGLISH CORPORA

In order to evaluate the appropriateness of the *EmotiBlog* annotation scheme and to prove that the fine-grained level it aims at has a positive impact on the performance of the systems employing it as training, we performed several experiments.

Given that a) *EmotiBlog* contains annotations for individual words, as well as for multi-word expressions and at a sentence level, and b) they are labeled with polarity, but also emotion, our experiments show how the annotated elements can be used as training for the opinion mining and polarity classification task, as well as for emotion detection. Moreover, taking into consideration the fact that *EmotiBlog* labels the intensity level of the annotated elements; we performed a brief experiment on determining the sentiment intensity, measured on a three-level scale: low, medium and high.

In order to perform these three different evaluations, we chose three different corpora. The first one is a collection of quotes (reported speech) from newspaper articles presented by Balahur et al. (Balahur et al., 2010d), enriched with the manual fine-grained annotation of *EmotiBlog*; the second one is the collection of newspaper titles in the test set of the SemEval 2007 task number 14 – Affective Text. Finally, the third one is a corpus of self-reported emotional response – ISEAR (Scherer and Walbott, 1999). The intensity classification task is evaluated only on the second corpus, given that it is the only one in which scores between -100 and 0 and 0 and 100, respectively, are given for the polarity of the titles.

CREATION OF TRAINING MODELS

For the OM and polarity classification task, we first extracted the Named Entities contained in the annotations using Lingpipe and united through a "_" all the tokens pertaining to the NE. All the annotations of punctuation signs that had a specific meaning together were also united under a single punctuation sign. Subsequently, we processed the annotated data using Minipar. We compute, for each word in a sentence, a series of features (some of these features are used by Choi et al. (2005)):

- the part of speech (POS);
- capitalization (if all letters are in capitals, if only the first letter is in capitals, and if it is a NE or not);
- opinionatedness/intensity/emotion - if the word is annotated as opinion word, its polarity, i.e. 1 and -1 if the word is positive or negative, respectively and 0 if it is not an opinion word, its intensity (1.2 or 3) and 0 if it is not a subjective word, its emotion (if it has, none otherwise);

- syntactic relatedness with other opinion word – if it is directly dependent of an opinion word or modifier (0 or 1), plus the polarity/intensity and emotion of this word (0 for all the components otherwise);
- role in 2-word, 3-word and 4-word annotations: opinionatedness, intensity and emotion of the other words contained in the annotation, direct dependency relations with them if they exist and 0 otherwise.

Finally, we add for each sentence as feature binary features for subjectivity and polarity, the value corresponding to the intensity of opinion and the general emotion. These feature vectors are fed into the Weka[31] SVM SMO machine learning algorithm and a model is created (EmotiBlog I). A second model (EmotiBlog II) is created by adding to the collection of single opinion and emotion words annotated in EmotiBlog, the Opinion Finder lexicon and the opinion words found in MicroWordNet, the General Inquirer resource and WordNet Affect.

EVALUATION RESULTS OF EMOTIBLOG-BASED MODELS ON TEST SETS

In order to evaluate the performance of the models extracted from the features of the annotations in *EmotiBlog*, we performed different tests. The first one regarded the evaluation of the polarity and intensity classification task using the *Emoitblog* I and II constructed models on two test sets – the JRC quotes collection and the SemEval 2007 Task Number 14 test set. Since the quotes often contain more than a sentence, we consider the polarity and intensity of the entire quote as the most frequent result in each class, corresponding to its constituent sentences. Also, given the fact that the SemEval Affective Text headlines were given intensity values between -100 and 100, we mapped the values contained in the Gold Standard of the task into three categories: [-100, -67] is high (value 3 in intensity) and negative (value -1 in polarity), [-66, 34] medium negative and [33, 1] is low negative. The values between [1 and 100] are mapped in the same manner to the positive category. 0 was considered objective, so containing the value 0 for intensity. The results are presented in Table 4.30 (the values I and II correspond to the models EmotiBlog I and EmotiBlog II):

Test Corpus	Evaluation type	Precision	Recall
JRC quotes I	Polarity	32.13	54.09
	Intensity	36.00	53.2
JRC quotes II	Polarity	36.4	51.00
	Intensity	38.7	57.81

[31] http://www.cs.waikato.ac.nz/ml/weka/

Test Corpus	Evaluation type	Precision	Recall
SemEval I	Polarity	38.57	51.3
	Intensity	37.39	50.9
SemEval II	Polarity	35.8	58.68
	Intensity	32.3	50.4

Table 4.30: Results for polarity and intensity classification using models built on EmotiBlog

The results presented in Table 4.30 show a significantly high improvement over the results obtained in the SemEval task in 2007. This is explainable, on the one hand, by the fact that systems performing the opinion task did not have at their disposal the lexical resources for opinion employed in the *EmotiBlog* II model, but also because of the fact that they did not use machine learning on a corpus comparable to *EmotiBlog* (as seen from the results obtained when using solely the *EmotiBlog* I corpus). Compared to the NTCIR 8 Multilingual Analysis Task this year, we obtained significant improvements in precision, with a recall that is comparable to most of the participating systems.

In the second experiment, we tested the performance of emotion classification using the two models built using EmotiBlog on the three corpora – JRC quotes, SemEval 2007 Task No.14 test set and the ISEAR corpus. The JRC quotes are labeled using EmotiBlog; however, the other two are labeled with a small set of emotions – 6 in the case of the SemEval data (joy, surprise, anger, fear, sadness, disgust) and 7 in ISEAR (joy, sadness, anger, fear, guilt, shame, disgust). Moreover, the SemEval data contains more than one emotion per title in the Gold Standard, therefore we consider as correct any of the classifications containing one of them. In order to unify the results and obtain comparable evaluations, we assessed the performance of the system using the alternative dimensional structures defined by Boldrini et al. (2009). The ones not overlapping with the category of any of the 8 different emotions in SemEval and ISEAR are considered as “Other” and are not included either in the training, nor test set. The results of the evaluation are presented in Table 4.31. Again, the values I and II correspond to the models EmotiBlog I and II. The “Emotions” category contains the following emotions: joy, sadness, anger, fear, guilt, shame, disgust, surprise.

Test corpus	Evaluation type	Precision	Recall
JRC quotes I	Emotions	24.7	15.08

Test corpus	**Evaluation type**	**Precision**	**Recall**
JRC quotes II	Emotions	33.65	18.98
SemEval I	Emotions	29.03	18.89
SemEval II	Emotions	32.98	18.45
ISEAR I	Emotions	22.31	15.01
ISEAR II	Emotions	25.62	17.83

Table 4.31: Results for emotion classification using the models built on EmotiBlog

The best results for emotion detection were obtained for the "anger" category, where the precision was around 35 percent, for a recall of 19 percent. The worst results obtained were for the ISEAR category of "shame", where precision was around 12 percent, with a recall of 15 percent. We believe this is due to the fact that the latter emotion is a combination of more complex affective states and it can be easily misclassified to other categories of emotion. Moreover, from the analysis performed on the errors, we realized that many of the affective phenomena presented were more explicit in the case of texts expressing strong emotions such as "joy" and "anger", and were mostly related to common-sense interpretations of the facts presented in the weaker ones.

As it can be seen in Table 4.30, results for the texts pertaining to the news category obtain better results, most of all news titles. This is due to the fact that such texts, although they contain a few words, have a more direct and stronger emotional charge than direct speech (which may be biased by the need to be diplomatic, find the best suited words etc.). Finally, the error analysis showed that emotion that is directly reported by the persons experiencing is more "hidden", in the use of words carrying special signification or related to general human experience. This fact makes emotion detection in such texts a harder task. Nevertheless, the results in all corpora are comparable, showing that the approach is robust enough to handle different text types. All in all, the results obtained using the fine and coarse-grained annotations in *EmotiBlog* increased the performance of emotion detection as compared to the systems in the SemEval competition.

DISCUSSION ON THE OVERALL RESULTS

From the results obtained, we can see that this approach combining the features extracted from the EmotiBlog fine and coarse-grained annotations helps to balance between the results obtained for precision and recall. The impact of using additional resources that contain opinion words is that of increasing the recall of the system, at the cost of a slight drop in precision, which shows that the approach is robust enough so that additional knowledge sources can be added. Although the corpus is small, the results obtained show that the phenomena captured by the approach are relevant to the opinion mining task, not only for the blog sphere, but also for other types of text (newspaper articles, self-reported affect).

Another advantage of EmotiBlog is the fact that it contains texts in three languages: English, Spanish and Italian. That is why, in the following experiments, we will test the usability of the resource in a second language, namely, Spanish.

4.4.4. USING EMOTIBLOG TO DETECT AND CLASSIFY SENTIMENT IN SPANISH TEXTS

Our annotation model includes word/ phrase/ text levels of annotation. It is thus useful for constructing similarity models for the training of machine learning algorithms working with different values of n-grams, as well as sentences as a whole.

EmotiBlog can be used to extrinsically evaluate systems mining opinions. Moreover, our annotation scheme can be used either for basic tasks of sentiment polarity classification, as well as emotion detection, either on very fine-grained categories, as well as psychology-based emotion classes. Furthermore, most work done in opinion mining only concentrated on classifying polarity of sentiments into positive or negative. Thus, another contribution our work brings is the classification according to three categories: positive, negative and neutral. The last category is made up of both opinionated sentences in which there is no clear indication of approval or disapproval of an idea, as well as objective sentences. Thirdly, our research focused both on the sentiment polarity classification, as well as the ratio between computational costs versus performance. Since our final goal is to employ the system in a real life scenario where it would adequately respond to user opinionated input with multimedia feedback, we must be aware of the advantages and disadvantages the use of each resource has on the system. Therefore, another contribution we bring is the assessment of performance versus time ratios. Last, but not least, we contribute to the research in the field by proposing and evaluating a method for sentiment polarity classification, based on n-gram and phrase similarity features used with machine learning. We evaluate the method both “intrinsically” – by cross-fold validation of the subjective and phrases in the corpus, as well as “extrinsically”, using a corpora of negative, positive and neutral opinions on recycling. The system was trained and tested using the annotation in the created corpus. However, given a different domain, it can be equally used, given that training examples are available.

The main aim of this experiment, described by Balahur et al. (Balahur et al., 2009b) is to obtain a system able to mine opinion from user input in real time and, according to the inferred information on the polarity of the sentiment, offer them corresponding feedback. The topic of our opinion mining experiment is “recycling”: the computer asks about a person’s opinion on recycling and then the user answers this question generating a sentence with emotion that can be of different intensity levels. The system reacts to the user input with some messages or faces that correspond to the reactions for the user’s feedback.

For the task at hand we employ annotations from texts on the Kyoto protocol pertaining to the EmotiBlog corpus. We use the annotated elements to train our opinion mining system and then classify new sentences that are on the topic “recycling” (to which some vocabulary similarity can be found, since they both topics refer to environmental issues). For the task at hand, we manually created a set of 150 sentences on recycling, 50 for each of the positive, negative and neutral categories. The first experiment carried out aimed at proving that the corpus is a valid resource and we can use the annotations for the training of our opinion mining system. For this assessment, we use the same methodology we will further employ to mine opinions from user input.

CROSS-FOLD EVALUATION OF THE ANNOTATION

As a result of the annotation, we obtained 1647 subjective phrases and 1336 objective ones. Our agreement was 0.59 for subjective phrases and 0.745 for the objective one.

Further on, we will consider for our tests only the sentences upon which we agreed and the phrases whose annotation length was above four tokens of the type noun, verb, adverb or adjective. For the cross-validation of the corpus, each of the sentences is POS-Tagged and lemmatized using FreeLing[32]. Further on, we represent each sentence as a feature vector, whose components are unigram features containing the positive and respectively negative categories of nouns, verbs, adverbs, adjectives, prepositions and punctuation signs (having “1” in the corresponding position of the feature vector for the words contained and “0” otherwise), the number of bigrams/ trigrams and 4-grams overlapping with each of the phrases we have annotated as positive and negative or objective, respectively and finally the overall similarity given by the number of overlapping words with each of the positive and negative or objective phrases from the corpus, normalized by the length of the given phrase. We test out method in two steps: first of all the classification of sentences among subjective and objective, for which the vectors contain as final values “subjective” or “objective” and second of all the classification of subjective sentences into positive and negative, for which case the

[32] http://www.lsi.upc.edu/~nlp/freeling/

classification vectors contain the values “positive and “negative”. We perform a ten-fold cross validation of the corpus for each of the two steps. The results are presented in Table 4.32, in terms of precision, recall and kappa.

	Precision	**Recall**	**Kappa**
Subjective	0.988	0.6	0.43
Objective	0.68	0.89	0.43
Positive	0.799	0.511	0.68
Negative	0.892	0.969	0.68

Table 4.32: Results of ten-fold corpus cross-validation

CLASSIFICATION OF NEW EXAMPLES

The second experiment we performed concerned the assessment of the system’s performance as far as the sentiment polarity classification of the sentences on recycling is concerned. Here, the challenge we are facing is the possible difference in the lexicon used. However, we assume the affective vocabulary to be approximately the same. We will consider as subjective neutral the sentences that are more similar to the “objective” statements in the training corpus. We will also test the importance of the fine-grained annotations on the classification performance.

Therefore, we will consider two scenarios: one in which we use the unigrams given by the annotated nouns, adjectives, verbs and adverbs from the corpus that we find in the phrases to be annotated; in the second scenario, we will only use the n-gram overlaps with n greater than 2 and the overall sentence similarity. Thus, we can obtain an overall evaluation of the importance of detecting also single words that have an affective charge.

CLASSIFICATION USING ALL N-GRAM FEATURES

For this first classification of our test data, we first run FreeLing on the set of positive, negative and neutral sentences on recycling in order to lemmatize and tag each word on part of speech. We then represent each sentence as a feature vector, in the same manner as in the first conducted experiment. Further on, we conduct two experiments on this data. The first one aims at training an SVM classifier on the corpus phrases pertaining to the “subjective” versus “objective” categories and test it on the statements on the recycling topic pertaining to the positive or negative versus neutral categories. The second experiment consists in classifying the instances according to three polarity classes: positive, negative and neutral. The results of the two experiments are summarized in Table 4.33.

	Precision	**Recall**	**Kappa**
Subjective	0.977	0.619	0.409
Objective	0.44	0.95	0.409
Positive	0.881	0.769	0.88
Negative	0.92	0.96	0.88

Table 4.33: Classification results using n-grams

As we can notice from the results, using the annotated elements, it is easier to distinguish the subjective sentences, due to the fact that we train on subjective n-grams. As far as the positive, negative and neutral classification is concerned, the results are both high, as well as balanced, proving the correctness of our approach.

CLASSIFICATION USING N-GRAMS, N>2.

In this experiment, we test the importance of annotating affect in texts at the token level. From our blog corpus, we have a large number of nouns, verbs, adverbs and adjective, annotated as positive or negative and at the emotion level. We used these words at the time of classifying examples using n-grams, with n ranging from 1 to 4 (in 5.2.1). To test their importance, we removed the vector components accounting for their presence in the feature vectors and re-classified, both at the level of objective versus subjective, as well as at the positive, negative, neutral level. In the table below, we can see the results obtained.

	Precision	**Recall**	**Kappa**
Subjective	0.93	0.60	0.43
Objective	0.43	0.7	0.43
Positive	0.83	0.64	0.85
Negative	0.90	0.91	0.85
Neutral	0.90	0.96	0.85

Table 4.34: Classification results using n-grams, n>2

As we can see, removing single words with their associated polarities from the training data resulted in lower scores. Therefore, fine-grained annotation helps at the time of training the opinion mining system and is well-worth the effort.

4.5. CONCLUSIONS ON THE PROPOSED METHODS FOR SENTIMENT ANALYSIS

In this chapter, we presented different methods and resources we built for the task of sentiment analysis in different text types.

We started by presenting methods to tackle the task of feature-based opinion mining and summarization, applied to product reviews. We have analyzed the peculiarities of this task and identified the weak points of existing research. We proposed and evaluated different methods to overcome the identified challenges, among which the most important were the discovery of indirectly mentioned features and the computation of the polarity of opinions in a manner that is feature-dependent. Subsequently, we proposed a unified model for sentiment annotation for this type of text, able to capture the important phenomena that we had identified – different types of sentiment expressions, feature mentioning and span of text expressing a specific opinion.

Further on, we explored different methods to tackle sentiment analysis from newspaper articles. After the initial experiments, we analyzed the reasons for the low performance obtained and redefined the task, taking into account the peculiarities of this textual genre. We created an annotation model and labeled two different corpora of newspaper article quotations, in English and German. After redefining the task and delimiting the scope of the sentiment analysis process to quotations – small text snippets containing direct speech, whose source and target are previously known-, the annotation agreement rose significantly. Additionally, improving the definition of the task made it possible to implement automatic processing methods that are appropriate for the task and significantly improve the performance of the sentiment analysis system we had designed. In the view of applying sentiment analysis to different types of texts, in which objective content is highly mixed with subjective one and where the sources and targets of opinions are multiple, we have proposed different general methods for sentiment analysis, which we applied to political debates.

The results of this latter experiment motivated us to analyze the requirements of a general labeling scheme for the task of sentiment analysis, which can be used to capture all relevant phenomena in sentiment expression.

To this aim, Boldrini et al. (2009) defined EmotiBlog, an annotation scheme that is able to capture, at a fine-grained level, all linguistic phenomena related to sentiment expression in text. The subsequent experiments have shown that this model is appropriate for the training of machine learning models for the task of sentiment analysis in different textual genres, in both languages in which experiments have been carried out using it – English and Spanish.

In this chapter, we have only concentrated on the task of sentiment analysis as a standalone challenge, omitting the steps required in order to obtain the texts on which the sentiment analysis methods were applied. In a real scenario, however, automatically detecting the opinion expressed in a text is not the first task to be performed. Additionally, in many of the cases, the results obtained after automatically processing texts to determine the sentiment they contain still pose many problems in terms of volume. Thus, even if the sentiment is determined automatically, one may still require a summarization component, in order to further reduce the quantity of information so that it is usable by a person.

Therefore, real-world applications that contain an opinion mining component must also contemplate the integration with other NLP systems. In the next chapter, we describe the challenges faced when integrating opinion mining systems with other NLP technologies, such as information retrieval (IR), question answering (QA) and text summarization (SUM). We first propose different methods to employ sentiment analysis in the context of NLP systems that are traditionally used for the analysis of factual data. Starting from the low results obtained in these preliminary efforts, we describe and evaluate new methods to tackle traditional NLP tasks (such as IR, QA and SUM) in the context of opinionated content. Our objective is to redefine these tasks, so that the phenomena encountered in opinionated texts are taken into account at every step of the processing, thus improving the overall performance of final systems.

CHAPTER 5. APPLICATIONS OF SENTIMENT ANALYSIS

Motto: *"He who molds public opinion makes statues and decisions possible or impossible to make." (Abraham Lincoln)*

5.1. INTRODUCTION

In the previous chapter, we presented different methods to perform sentiment analysis from a variety of text types. We have shown what the challenges related to each of these genres are and how the task of opinion mining can be tackled in each of them.

Nevertheless, real-world applications of sentiment analysis often require more than an opinion mining component. On the one hand, an application should allow a user to query about opinions, in which case the documents in which these opinions appear have to be retrieved. In more user-friendly applications, the users can be given the option to formulate the query into a question in natural language.

Therefore, question answering techniques must be applied in order to determine the information required by the user and subsequently retrieve and analyze it. On the other hand, opinion mining offers mechanisms to automatically detect and classify sentiments in texts, overcoming the issue given by the high volume of such information present on the Internet. However, in many cases, even the result of the opinion processing by an automatic system still contains large quantities of information, which are still difficult to deal with manually. For example, for questions such as *"Why do people like George Clooney?"* we can find thousands of answers on the Web. Therefore, finding the relevant opinions expressed on George Clooney, classifying them and filtering only the positive opinions is not helpful enough for the user. He/she will still have to sift through thousands of texts snippets, containing relevant, but also much redundant information. Moreover, when following the comments on a topic posted on a blog, for example, finding the arguments given in favor and against the given topic might not be sufficient to a real user. He/she might find the information truly useful only if it is structured and has no redundant pieces of information. Therefore, apart from analyzing the opinion in text, a real-world application for sentiment analysis could also contain a summarization component.

The aim of the work presented in this chapter is to apply the different opinion mining resources, tools and approaches to other tasks within NLP. The objective was, on the one hand, to evaluate the performance of our approaches, and, on the other, to test the requirements and extra needs of an opinion mining system in the context of larger applications. In this chapter, we present the research we carried out in order to test the manner in which opinion mining can be best combined with information retrieval (IR), question answering (QA) and summarization (SUM), in order to create a useful, real-life, end-to-end system for opinion analysis in text. Our initial efforts concentrated on applying already-existing IR, QA and SUM systems in tandem with our sentiment analysis systems. Subsequently, having realized that directly applying systems that were designed to deal with factual data in the context of opinionated text led to low results, we proposed new methods to tackle IR, QA and SUM in a manner that is appropriate in the context of subjective texts.

In this chapter, we present the methods and improvements achieved in Opinion Question Answering and Opinion Summarization.

5.2. OPINION QUESTION ANSWERING AND SUMMARIZATION

While techniques to retrieve objective information have been widely studied, implemented and evaluated, opinion-related tasks still represent an important challenge. As a consequence, the aim of this section of research is to study, implement and evaluate appropriate methods for the task of Question Answering (QA) in the context of opinion treatment.

The experience in the TAC 2008 Opinion Pilot competition (Balahur et al., 2008), as well as the post-competition experiments (Lloret et al., 2009), motivated us to continue the research we started with the participation in the TAC competition with the study of different aspects of opinion Question Answering and opinion summarization, such as classification of fact versus opinion questions (Balahur et al., 2009c), defining opinion answer types and methods to retrieve answers to opinion questions (Balahur et al., 2009a; Balahur et al., 2009d; Balahur et al., 2009h; Balahur et al., 2010a; Balahur et al., 2010b; Balahur et al., 2009e) or opinion-driven summarization in the context of threads in blogs (Balahur et al., 2010e; Balahur et al., 2009g; Balahur et al., 2009i; Kabadjov et al., 2009).

5.2.1. PARTICIPATION TO THE TAC 2008 OPINION PILOT TASK

The Text Analysis Conference 2008 edition (TAC 2008) contained three tasks: Question Answering, Recognizing Textual Entailment and Summarization. The Summarization track included two tasks: the Update Summarization and the Opinion Summarization Pilot task. We participated in the Opinion Summarization Pilot task within the Summarization track, with the aim of measuring the performance of our opinion mining system (Balahur and Montoyo, 2008c; Balahur and Montoyo, 2008d). However, given the fact that the referenced system was built to deal only with opinions on products and their features, in the end we changed our approach to a more general one, which is similar to the one used to determine opinions from political debates (Balahur et al., 2009e).

The Opinion Summarization Pilot task consisted in generating summaries from blogs snippets that were the answers to a set of opinion questions. The participants were given a set of blogs from the Blog06 collection and a set of "squishy list" (opinion) questions from the Question Answering track, and had as task to produce a summary of the blog snippets that answered these questions. There were 25 targets, and on each of them one or two questions were formulated. All these questions concerned the attitude held by specified sources on the targets given. For example, for the target "George Clooney", the two questions asked were *"Why do people like George Clooney?"* and *"Why do people dislike George Clooney?"*). Additionally, a set of text snippets were also provided, which contained the answers to the questions. These snippets were selected from the answers given by systems participating in the Question Answering track, and opinion summarization systems could either use them or choose to perform themselves the retrieval of the answers to the questions in the corresponding blogs.

Within our participation in the Opinion Summarization Pilot task, we used two different methods for opinion mining and summarization. The two approaches suggested were different only as far as the use of the optional text snippets provided by the TAC organization was concerned. Our first approach (the Snippet-driven Approach) used these snippets, whereas the second one (Blog-driven Approach) found the answers directly in the corresponding blogs.

In the first phase, we processed the questions, in order to determine a set of attributes that will further help us find and filter the answers. The process is described in Figure 5.1. In order to extract the topic and determine the question polarity, we define question patterns. These patterns take into consideration the interrogation formula and extract the opinion words (nouns, verbs, adverbs, adjectives and their determiners). The opinion words are then classified in order to determine the polarity of the question, using the WordNet Affect emotion lists, the emotion triggers resource (Balahur and Montoyo, 2008), a list of four attitudes that we built, containing the verbs, nouns, adjectives and adverbs for the categories of criticism, support, admiration and rejection and two categories of value words (good and bad) taken from the opinion mining system proposed by Balahur and Montoyo (Balahur and Montoyo, 2008c).

Examples of rules for the interrogation formula "What reasons" are:

1. *What reason(s) (.*?) for (not) (affect verb + ing) (.*?)?*
2. *What reason(s) (.*?) for (lack of) (affect noun) (.*?)?*
3. *What reason(s) (.*?) for (affect adjective ||positive ||negative) opinions (.*?)?*

Using these extraction patterns, we identified the nouns, verbs, adjectives etc. that gave an indication of the “question polarity”[33]. Further on, these indicators were classified according to the affect lists mentioned above. The keywords of the question are determined by eliminating the stop words. At the end of the question processing stage, we obtain, on the one hand, the reformulation patterns (that are eventually used to link and give coherence to the final summaries) and, on the other hand, the question focus, keywords and the question polarity. Depending on the focus/topic and polarity identified for each question, a decision on the further processing of the snippet was made, using the following rules:

1. If there is only one question made on the topic, determining its polarity is sufficient for making the correspondence between the question and the snippets retrieved; the retrieved snippet must simply obey the criteria that it has the same polarity as the question.
2. If there are two questions made on the topic and each of the questions has a different polarity, the correspondence between the question and the answer snippets can simply be done by classifying the snippets retrieved according to their polarity.
3. If there are two questions that have different focus but different polarities, the correspondence between the questions and the answer snippets is done using the classification of the answer snippets according to focus and polarity.
4. If there are two questions that have the same focus and the same polarity, the correspondence between the questions and the answer snippets is done using the order of appearance of the entities in focus, both in the question and in the possible answer snippet retrieved, simultaneously with the verification that the intended polarity of the answer snippet is the same as that of the question.

The categorization of questions into these four classes is decisive at the time of making the question - answer snippet correspondence, in the snippet/blog phrase processing stage. Details on these issues are given in what follows and in Figure 5.1.

[33] Later in this chapter, as we will redefine the task of question answering in the context of opinions, we will refer to this concept as “Expected Polarity Type” (EPT). Although in this first experiment, we called it “question polarity”, the new EPT term was necessary, as “question polarity” entails that the question in itself has some sort of orientation, when in fact it is the polarity of the expected answer that is actually computed at this stage.

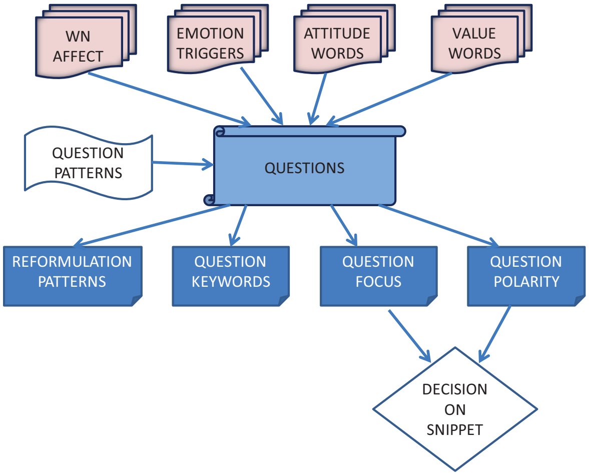


Figure 5.1: Question Processing Stage

In the first approximation, the given answer-snippets constitute the basis for looking up the phrases these snippets were extracted from in the blogs collection. In the second approximation, we use the question keywords to determine the phrases from the blogs that could constitute answers to the questions. Further on, the blog original phrases or the blog retrieved phrases, respectively, are classified according to polarity, using the vector similarity with the set of vectors consisting of three distinct subsets.

The first subset of vectors is built from the phrases in the ISEAR corpus (without stop words), one vector per statement and having the phrases classified according to the emotion it described. The second subset of vectors is built according to the WN Affect list of words in the joy, anger, sadness and fear. The third subset consists of the vectors of emotions from the emotion triggers resource, on each of the 4 categories that were considered also for building the vectors from WordNet Affect.

For each of the blog phrase, we compute the similarity with all the vectors. Further on, each of the emotions is assigned a polarity - emotions from the fear, anger, sadness, disgust, shame, guilt categories are assigned a negative polarity - and emotions from the joy and surprise categories we considered as positive. The final polarity score was computed as sum of the scores obtained in each of the vector similarity computations. The higher of the two scores - positive or negative - is considered as being the snippet polarity. In the second approximation, we also perform a sorting of the phrases retrieved, in descending order, according to their polarity scores. This is helpful at the time of building the final summary, whose length must not surpass a given limit (in this case, the organizers set the limit to 7000 characters).

In the final phrases used in creating the summary we added, for coherence reasons, the reformulation patterns deduced using the question structure. Taking into consideration the number of characters limitation, we only included in the summary the phrases with high positive scores and those with high negative scores, completed with the reformulation patterns, until reaching the imposed character limit. Thus, the score given by the sentiment analysis system constituted the main criteria for selecting the relevant information for the summaries and the F-measure score reflects the quality of the opinion mining process.

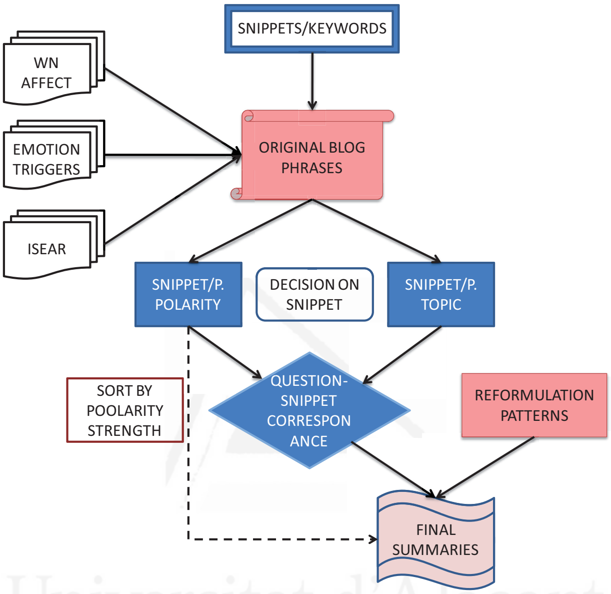


Figure 5.2: The snippet/ blog phrase processing stage

In the first approach, we used the snippets provided by the organizers. We automatically analyzed them in order to determine the focus and polarity in each of the given questions for a topic, determine the associated given answer snippet (by computing the snippet's polarity and focus), locate the whole sentence's snippet within the corresponding blog, and finally use patterns of reformulation from the questions' structure to bind together the snippets for the same polarity and focus to produce the final summary. Having computed the polarity for the questions and text snippets, and having set out the final set of sentences to produce the summary with their focus, we bound each sentence to its corresponding question, and we grouped all sentences which were related to the same question together, so that we could generate the language for this group, according to the patterns of reformulation that were created for each question. Finally, the speech style was changed to an impersonal one, in order to avoid the presence of directly expressed opinion sentences. A POS-tagger tool (TreeTagger) was used to identify third person verbs and change them to a neutral style. A set of rules to identify pronouns was created, and they were also changed to the more general pronoun “they” and its corresponding forms, to avoid personal opinions.

The second approach had as starting point determining the focus, keywords, topic and polarity in each of the given questions. The processing of the question is similar to the one performed for the first approximation. Starting from the focus, keywords and topic of the question, we sought sentences in the blog collection (previously processed as described in the first approximation) that could constitute possible answers to the questions, according to their similarity to the latter. The similarity score was computed with Pedersen’s Text Similarity Package[34]. The snippets thus determined underwent dependency parsing with Minipar and only the sentences which contained subject and predicate were kept, thus ensuring the elimination of some of the present “noise” (such as section titles, dates, times etc.). The remaining snippets were classified according to their polarity, using the similarity score with respect to the described emotion vectors. The direct language style was changed to indirect speech style. The reformulation patterns that were deduced using the questions’ structure were added to bind together the snippets and produce the final summary, concatenating the snippets with the added reformulations. Since the final length of the summary could easily overpass the imposed limit, we sorted the snippets using their polarity strength (the higher the polarity score - be it positive or negative the higher the rank of the snippet), and included the reformulated snippets in descending order until the final limit was reached.

EVALUATION OF THE APPROACH IN THE TAC2008 COMPETITION

45 runs were submitted by 19 teams for evaluation in the TAC 2008 Opinion Pilot task. Each team was allowed to submit up to three runs, but finally, due to the difficulty involved in the evaluation of such a task, only the first two runs of each team was evaluated, leading to 36 runs being evaluated. Table 5.1 shows the final results obtained by the first two runs we submitted for evaluation in the TAC 2008 Opinion Pilot and the rank they had compared to the other participating systems. The column numbers stand for the following information:

1. summarizerID (our Run 1 had summarizerID 8 and Run 2 had summarizerID 34)
2. Run type: "manual" or "automatic"
3. Did the run use the answer snippets provided by NIST: "Yes" or "No"
4. Average pyramid F-score (Beta=1), averaged over 22 summaries
5. Average score for Grammaticality

[34] http://wn-similarity.sourceforge.net/

6. Average score for Non-redundancy
7. Average score for Structure/Coherence (including focus and referential clarity)
8. Average score for Overall fluency/readability
9. Average score for Overall responsiveness

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
8	automatic	Yes	0.357	4.727	5.364	3.409	3.636	5.045
34	automatic	No	0.155	3.545	4.364	3.091	2.636	2.227

Table 5.1: Evaluation results in the TAC 2008 competition

Further on, we will present the system performances with respect to all other teams, first as an overall classification (Table 5.2) and secondly, taking into consideration whether or not the run used the optional answer snippets provided by NIST (Table 4).

In Table 5.2, the numbers in columns 4, 5, 6, 7, 8 and 9 correspond to the position within the 36 evaluated submissions. In Table 5.3, the numbers in columns 4, 5, 6, 7, 8 and 9 correspond to the position within the 17 submissions that used the given optional answer snippets (in case of Run 1) and the position within the 19 submissions evaluated that did not use the provided answer snippets.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
8	automatic	Yes	7	8	28	4	16	5
34	automatic	No	23	36	36	13	36	28

Table 5.2: Classification results (overall comparison)

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
8	automatic	Yes	7	15	14	2	11	5
34	automatic	No	9	19	19	6	19	14

Table 5.3: Classification results (comparison with systems using/not using answer snippets)

As it can be noticed from the results table, our system performed well regarding Precision and Recall, the first run being classified 7th among the 36 evaluated runs as far as F-measure. As far as the structure and coherence are concerned, the results were also good, placing Run 1 in 4th position among the 36 evaluated runs. Also worth mentioning is the good performance obtained as far as the overall responsiveness is concerned, where Run 1 ranked 5th among the 36. When comparing our approaches separately, in both cases, they did not perform very well with respect of the non-redundancy criterion, nor the grammaticality one. An interesting thing that is worth mentioning as far as the results obtained are concerned, is that the use of reformulation patterns, in order to generated sentences for completing the summaries, has been appropriate, leading to very good rankings according to the structure/coherence criterion. However, due to the low results obtained as far as the redundancy and grammaticality criteria are concerned, we decided to test different methods to overcome these issues.

5.2.2. POST TAC 2008 COMPETITION EXPERIMENTS AND RESULTS

As we have seen from the results obtained in the TAC 2008 Opinion Pilot competition results, when using the optional snippets, the main problem to address is to remove redundancy. Many of the text snippets provided by the organizers repeat the same arguments for the same target, although with slightly different words. In order to create non-redundant summaries (and in this manner, also be able to improve on the F-score, as eliminating useless information allows to include more informative text snippets in the final summary, without surpassing the imposed 7000 characters limit), we must determine which of the snippets represent better an idea for the final summary and remove all the other snippets that express the same thought. To this aim, in Lloret et al. (2009), we performed several experiments, involving the use of a *summarization system* (Lloret et al., 2008) and a *textual entailment system* (Iftene and Balahur-Dobrescu, 2007).

Several participants in the TAC 2008 edition performed the opinion summarization task by using generic summarization systems. Their results were significantly lower as far as F-measure or overall responsiveness was concerned, as their systems were specifically designed to perform summarization on factual data. Therefore, we use the summarization system and the textual entailment system prior to classifying opinion and selecting the highest scored snippets for generating the final summaries. The generic summarization system employed was the one described by Lloret et al. (2008). The main idea of this proposal is to score sentences of a document with regard to the word frequency count (WF)[35], which can be combined with a Textual Entailment (TE) module. The TE system created by Iftene and Balahur-Dobrescu (2007) contemplates the entailment at different levels – lexical, syntactic and semantic and is based on the tree-edit distance algorithm (Kouleykov and Magnini, 2006).

Textual entailment (Dagan et al., 2006) is the task of determining whether the meaning of one text snippet (the hypothesis, H) can be inferred by another one (the text, T). Systems implementing this task have been competing since 2005 in the

[35] This principle states that the more times a word appears in a document, the more relevant the sentences that contain this word are.

“Recognising Textual Entailment” (RTE) Challenges (RTE)[36]. The following examples extracted from the development corpus provided by the 3rd RTE Challenge show a true and false entailment relation between two text snippets:

Pair id=50 (entailment = true)
T: “Edison decided to call “his” invention the Kinetoscope, combining the Greek root words “kineto”(movement), and “scopos” (“to view”).”
H: “Edison invented the Kinetoscope.”

Pair id=18 (entailment = false)
T: “Gastrointestinal bleeding can happen as an adverse effect of non-steroidal anti-inflammatory drugs such as aspirin or ibuprofen.”
H: “Aspirin prevents gastrointestinal bleeding.”

Although the first approach suggested for opinion summarization obtained much better results in the evaluation than the second one, we decided to run the generic summarization system by Lloret et al. (2008) over both approaches, with and without applying TE, to provide a more extent analysis and conclusions. After preprocessing the blogs and having all the possible candidate sentences grouped together, we considered these as the input for the generic summarizer.

The goal of these experiments was to determine whether the techniques used for a generic summarizer would have a positive influence in selecting the main relevant information to become part of the final summary. We re-evaluated the summaries generated by the generic system following the nuggets list provided by the TAC 2008 organization, and counting manually the number of nuggets that were covered in the summaries. This was a tedious task, but it could not be automatically performed because of the fact that many of the provided nuggets were not found in the original blog collection. After the manual matching of nuggets and sentences, we computed the average Recall, Precision and F-measure (Beta =1) in the same way as in the TAC 2008 was done, according to the number and weight of the nuggets that were also covered in the summary. Each nugget had a weight ranging from 0 to 1 reflecting its importance, and it was counted only once, even though the information was repeated within the summary. The average for each value was calculated taking into account the results for all the summaries in each approach. Table 4 points out the results for all the approaches reported. We have also considered the results derived from our participation in the TAC 2008 conference

[36] The RTE Challenges have been organized between 2005 and 2007 by the Pascal Network of Excellence and since 2008, by the National Institute for Standards and Technology, within the Text Analysis Conference (TAC).

(OpSum-1 and OpSum-2), in order to analyze whether they have been improved or not.

System	Recall	Precision	F-measure
OpSum-1	0.592	0.272	0.357
OpSum-2	0.251	0.141	0.155
WF-1	**0.705**	0.392	0.486
TE+WF -1	0.684	**0.630**	**0.639**
WF -2	0.322	0.234	0.241
TE+WF-2	0.292	0.282	0.262

Table 5.4: Comparison of the results obtained in the competition versus the results obtained applying the summarization system proposed by Lloret et al. (2008)

From these results it can be stated that the TE module in conjunction with the WF counts, have been very appropriate in selecting the most important information of a document. Although it can be thought that applying TE can remove some meaningful sentences which contained important information, results show the opposite. It benefits the Precision value, because a shorter summary contains greater ratio of relevant information. On the other hand, taking into consideration the F-measure value only, it can be seen that the approach combining TE and WF, for the sentences in the first approach, significantly improved the best F-measure result among the participants of TAC 2008, increasing its performance by 20% (with respect to WF only), and improving by approximately 80% with respect to our first approach submitted to TAC 2008.

However, a simple generic summarization system like the one we have used here is not enough to produce opinion oriented summaries, since semantic coherence given by the grouping of positive and negative opinions is not taken into account. Therefore, simply applying the summarization system does not yield the desired results in terms of opinionated content quality. Hence the opinion classification stage must be added in the same manner as used in the competition and combined appropriately with a redundancy-removing method.

Motivated by this first set of experiments, in a second approach, we wanted to test how much of the redundant information would be possible to remove by using a Textual Entailment system similar to the one proposed by Iftene and Balahur-Dobrescu (2007), without it affecting the quality of the remaining data. As input for the TE system, we considered the snippets retrieved from the original blog posts. We applied the entailment verification on each of the possible pairs, taking in turn all snippets as Text and Hypothesis with all other snippets as Hypothesis and Text, respectively. Thus, as output, we obtained the list of snippets from which we eliminated those that are entailed by any of the other snippets.

System	F-Measure
Best system	0.534
Second best system	0.490
OpSum-1 + TE	0.530
OpSum-1	0.357

Table 5.5. Comparison of the best scoring systems in TAC 2008 and the DLSIUAES team's improved system

Table 5.5 shows that applying TE before generating the final summary leads to very good results increasing the F-measure by 48.50% with respect to the original first approach. Moreover, it can be seen form Table 5.5 that our improved approach would have ranked in the second place among all the participants, regarding F-measure, maintaining the linguistic quality level, with which our approach ranked high in the TAC 2008 Opinion Pilot competition. The main problem with this approach is the long processing time. We can apply Textual Entailment in the manner described within the generic summarization system presented, successively testing the relation as "Snippet1 entails Snippet2?", "Snippet1+Snippet2 entails Snippet3?" and so on. The problem then becomes the fact that this approach is random, since different snippets come from different sources, so there is no order among them. Further on, we have seen that many problems arise from the fact that extracting information from blogs introduces a lot of noise. In many cases, we had examples such as:

"At 4:00 PM John said Starbucks coffee tastes great"
"John said Starbucks coffee tastes great, always get one when reading New York Times."

To the final summary, the important information that should be added is "Starbucks coffee tastes great". Our TE system contains a rule specifying that the existence or not of a Named Entity in the hypothesis and it not being mentioned in the text leads to the decision of "NO" entailment. For the example given, both snippets are maintained, although they contain the same data. Another issue to be addressed is the extra information contained in final summaries that is not scored as nugget. As we have seen from our data, much of this information is also valid and correctly answers the questions. Therefore, what methods can be employed to give more weight to some and penalize others automatically?

Regarding the grammaticality criteria, once we had a summary generated we used the module Language Tool[37] as a post-processing step. The errors that we needed correcting included the number matching between nouns and determiners as well as among subject and predicate, upper case for sentence start, repeated words or punctuation marks and lack of punctuation marks. The rules present in the module and that we "switched off", due to the fact that they produced more errors, were those concerning the limit in the number of consecutive nouns and the need for an article before a noun (since it always seemed to want to correct "Vista" for "the Vista" a.o.). We evaluated by observing the mistakes that the texts contained, and counting the number of remaining or introduced errors in the output. The results obtained can be seen in Table 5.6.

Problem	Rightly corrected	Wrongly corrected
Match S-P	90%	10%
Noun-det	75%	25%
Upper case	80%	20%
Repeated words	100%	0%
Repeated "."	80%	20%
Spelling mistakes	60%	40%
Unpaired ""/()	100%	0%

Table 5.6. Grammaticality analysis

The greatest problem encountered was the fact that bigrams are not detected and agreement is not made in cases in which the noun does not appear exactly after the determiner. All in all, using this module, the grammaticality of our texts was greatly improved. In our post-competition experiments, we showed that using a generic summarization system, we obtain 80% improvement over the results obtained in the competition, with coherence being maintained by using the same polarity classification mechanisms. Using redundancy removal with TE, as opposed to our initial polarity strength based sentence filtering improved the system performance by almost 50%. Finally, we showed that grammaticality can be checked and improved using an independent solution given by Language Tool.

Nevertheless, many of the components of this approach still require improvement. First of all, as we have seen, special methods must be applied in order to treat the questions, in order to determine additional elements (expected polarity of answer, source of opinion expected, target of opinion etc.). Additionally, as in the case of the initial QA task at TAC 2008, real-life systems must be able to distinguish between factual questions and questions that require an opinionated

[37] http://community.languagetool.org/

answer, in order to apply specific, appropriate methods to retrieve the correct answers. This motivated our subsequent research, which first concentrated on defining a method through which opinion questions could be differentiated from factual ones and subsequently on proposing and evaluating specific methods to tackle opinion question answering.

5.2.3. PROPOSAL OF A FRAMEWORK FOR OPINION QUESTION ANSWERING

Motivated by the experiments in the TAC 2008 competition and the subsequent efforts to improve the approaches proposed, we set out to determine what the particularities of opinion questions are, in order to be able to propose adequate methods to tackle them.

FACT VERSUS OPINION QUESTION CLASSIFICATION

Firstly, we propose an approach towards solving the problem of question classification in a mixed setting of opinion (multi-perspective) and fact question answering (Balahur et al., 2009c).

Moreover, we present a method for multi-perspective question answering based on the use of polarity classification and Textual Entailment. We show why the classification of questions is important at the time of answer retrieval and validation and what are the challenges we are faced with in answering opinion questions. The difficulties within this task are explained and solutions are proposed for each of the issues encountered. Finally, we evaluate the accuracy in question classification using the mixed test sets of the Answer Validation Exercise 2008, the TAC 2008 Opinion Pilot, the question set made on the OpQA corpus (Stoyanov et al., 2005), and the set of questions we created using our own annotated corpus *EmotiBlog* (Boldrini et al., 2009). The multi-perspective question answering method is evaluated using the OpQA corpus and *EmotiBlog*. We show the improvements brought by the approach in retrieving the correct answers. We discuss the implications of the method proposed in a Question Answering (QA) System that is able to retrieve answers for both fact and opinion questions.

Another relevant challenge is to determine the best way to look for the answers. In order to achieve this, we must first analyze the correct answers that are annotated in a corpus and check if it is difficult to retrieve them (or which of the methods and tools should be employed in order to take into account their peculiarities).

Using *EmotiBlog*, we label each answer adding three elements. The first one is the source, the second one is target and finally, we annotate the required polarity. Using these annotations, we will be able to detect the author of the sentence, the target of the sentiment and also the polarity of the expressed opinion.

Furthermore, we also have to find an effective method to check if an answer is correct. This is difficult for general opinion answers, because in most of the cases answers for the same question can be different but not for this reason incorrect. In fact, they could express different points of view about a subject. We could say that each of them is partially correct and none of them is totally correct.

After having retrieved the answers, answer validation is needed. We could perform the Answer Validation Exercise (AVE) for objective answers, but for opinion sentences we do not have at our disposal any similar tool. We would need for example a consistent collection of paraphrases for opinion but, to the best of our knowledge, such a collection does not exist.

In order to carry out our experiments, we created a small collection of questions in addition to the ones made by Stoyanov et al. (2005) for the OpQA corpus. We decided to use part of this collection because our corpus and the MPQA share the same topic, the Kyoto Protocol; as a consequence our corpus will probably contain the answer to the queries and our questions will also have answers in their corpus.

We used 8 questions in English, divided into two groups; the first one is composed by factoid ones and we consider them as objective queries. They usually require a single fact as answer and include questions such as: *What is the Kyoto Protocol?* or *When was the Kyoto Protocol ratified?* As we can deduce, they are asking for a definition and for a date.

In general, factoid queries ask for a name, date, location, time etc. and as a consequence, the most relevant aspect is that they need a univocal answer. Moreover, answers to factoid questions can be validated, using different techniques, such as Textual Entailment, as proposed in the Answer Validation Exercise (AVE)[38].

The second group is composed of opinion questions. They are different from the ones previously described. They are more complex in nature, as they contain additional elements, such as required polarity, and the answer is not univocal. Through their nature, opinion questions require more answers to be retrieved that are equally correct and valid (i.e. *What reasons do people have for liking MacDonald's?*). We will obtain a wide list of answers each of them will be correct.

Furthermore, there are opinion questions that could be interpreted as objective, but which are actually requiring the analysis of opinionated content in order to be

[38] http://nlp.uned.es/clef-qa/ave/

answered (e.g. *Are the Japanese unanimous in their opinion of Bush's position on the Kyoto Protocol?).* Usually, opinion queries can have more than one answer.

One of the possible answers found in the corpus under the form of *Japanese were unanimous in their opinion about Bush's position on the Kyoto protocol*, and it is equally possible that answering this question requires the analysis of many opinions on the same subject. They are evaluative or subjective because they express sentiments, feelings or opinions.

Table 5.7 presents the collection of questions we used in order to carry out our experiments.

Type	**Questions**
factoid	What is the Kyoto protocol about?
factoid	When was the Kyoto Protocol adopted?
factoid	Who is the president of Kiko Network?
factoid	What is the Kiko Network?
opinion	What is Bush's opinion about the Kyoto protocol?
opinion	What are people's feelings about Bush's decision?
opinion	What is the Japanese reaction to the Kyoto protocol?
opinion	What are people's opinions about Bush

Table 5.7: The set of questions proposed on the EmotiBlog corpus

As we can see in Table 5.7, we have a total of 8 questions, divided between factoid and opinion. After having created the collection of questions, we labeled the answers in our blog corpus using *EmotiBlog*, the annotation scheme we built for emotion detection in non-traditional textual genres, as for example blogs or forums. The granularity of the annotations done using this scheme could be an advantage if the model is designed in a simple, but effective way. Using EmotiBlog, we aim at capturing all relevant phenomena for the sentiment analysis task, in order to provide the opinion mining system the maximum amount of information.

In order to propose appropriate methods to tackle opinion questions, we must first determine what are the challenges associated, in comparison to factual questions. Therefore, we first studied the answers to opinion and factual questions in the OpQA corpus. In the case of factoid questions, we noticed that, in most of the cases, they are answerable in a straightforward manner, i.e. the information required is found in the corpus expressed in a direct manner. One problem we found was related to temporal expressions. For example, for the question *"When was the Kyoto Protocol ratified?"*, there are some answers that contain *1997*, but other answers are *"last November"*, or *"in the course of the year"*. As a consequence, we deduce that, in order to correctly interpret the answer, additional context must be used. Such extra details could be for example the date of the document. For the rest of factoid question we did not detect other relevant obstacles.

Regarding opinion queries, the analysis revealed a series of interesting issues. The first one is that some opinion questions could be interpreted as factoid. For example, if we have: *Are the Japanese unanimous in their opinion of Bush's position on the Kyoto Protocol?* This query could be answered with *yes/no*, but the information we really need is an opinion. Thus, the first thing to do should be to build up some patterns in order to detect if a question is about an opinion.

Another problem we detected analyzing OpQA answers is the lack of source of the opinion. If we ask: *How is Bush's decision not to ratify the Kyoto Protocol looked upon by Japan and other US allies?* This question is asking for two different points of view, but its corresponding answers are not clear. We detected for example: *We'd like to see further efforts on the part of the U.S.* or *dump*. As we can notice, we do not know who is expressing each of these opinions. This is another issue that should be solved. Therefore, when annotating the answers with EmotiBlog, we specify the target (the entity on which the opinion is expressed) and the source of the opinion.

In order to improve this task we could label our collection of questions using the Expected Answer Type (EAT) and our corpus with Named Entities (NE). On the one hand, the EAT would solve the problem of understanding each query type, and on the other hand NE labeling could improve the entity detection.

It is worth mentioning that opinion QA is a challenging and complex task. The first step we should be able to perform is to discriminate factoid versus opinion questions and, after having solved this first problem, we should try to find a way to expand each of our queries.

Finally, we should have a method to automatically verify the correctness of partial answers.

In order to study the task of fact versus opinion classification, we use on the one hand the sets provided within the AVE (Answer Validation Exercise) 2007 and 2008 development and test sets (fact questions) and the development and test sets in the TAC (Text Analysis Conference) 2008 Opinion Pilot, the questions made on the OpQA corpus (Stoyanov et al., 2005) and the questions we formulated on the *EmotiBlog* corpus. Finally, we gathered a total of 334 fact and 79 opinion questions. The first classification we performed was using the corpus annotations on the Kyoto protocol blog posts. For each of the questions, we computed the similarity it has with each of the objective versus subjective phrases annotated in the corpus and built vectors of the obtained scores. We then classified these vectors using ten-fold cross validation with SVM SMO.

The results obtained in this first phase are summarized in Table 5.8 (P denotes the Precision score, R the Recall, Acc. is the accuracy over the classified examples and Kappa measures the confidence of the classification).

Question type	P	R	Acc.	Kappa
Fact	0.891	0.955	0.86	0.52
Opinion	0.727	0.506		

Table 5.8: Fact versus opinion question classification results

As we can notice, opinion questions are more difficult to discriminate based only on the similarity to the objective/subjective vocabulary.

A second experiment we performed concerned taking into consideration the interrogation formula as clue for classifying the fact and opinion questions. We considered, aside from the similarity features employed in the first experiment, the type of interrogation formula employed. We had a list of interrogation formulas that were composed of one or two words at the beginning of the questions. They were: *who, what, why, when, how long, how much, where, which, what reason(s), what motive, are, is does, did, do, were, was and other.* The reason for considering also cases of interrogation formulas composed of two words was that many opinion and fact questions starting with "what" or "how" can only be discriminated on the basis of the word following these terms (e.g. "how long", "how much",   what reason(s)" etc.). All other formulations were considered as type "other".   The results obtained are summarized in Table 5.9.

Question type	P	R	Acc.	Kappa
Fact	0.923	0.987	0.92	0.73
Opinion	0.93	0.671		

Table 5.9: Fact versus opinion question classification using the interrogation formula

As we can notice from Table 5.9, when using the clue of the interrogation formula, the results improved substantially for the opinion question category classification. Further on, we performed a third experiment that had as aim to test the influence of the number of learning examples for each category on the classification results. As category of opinion questions category was much smaller than the number of examples for fact questions (79 as opposed to 334), we wanted to measure the performance in classification when the training sets were balanced for the two categories. We randomly selected 79 fact questions from the 334 initial ones and repeated the second experiment.

The results are summarized in Table 5.10.

Question type	P	R	Acc.	Kappa
Fact	0.908	0.747	0.83	0.67
Opinion	0.785	0.924		

Table 5.10 Fact versus opinion question classification using the interrogation formula and an equal number of training examples

In this case, we can see that recall significantly drops for the fact category and increases for the opinion examples. Precision for the fact category remains around the same values in all classification settings and drops for the opinion category when using fewer examples of fact questions. At an overall level, however, the results become more balanced. Therefore, a first conclusion that we can draw is that we need larger collections of opinion questions in order to better classify opinion and fact questions in a mixed setting.

Subsequently to this analysis of methods for fact versus opinion question classification, we annotated the answers corresponding to the set of questions we proposed, in the *EmotiBlog* corpus. Table 5.11 presents the questions and the number of answers annotated using the EmotiBlog annotation scheme.

Questions	Number of answers
What is Bush's opinion about the Kyoto protocol?	4
What are people's feelings about Bush's decision?	1
What is the Japanese reaction to the Kyoto Protocol	3
What are peoples' opinions about Bush?	4

Table 5.11 List of opinion questions and the number of corresponding answers annotated in the corpus

As we can see in table 5.11, we have labeled different answers for each opinion questions. Each of these questions is different in nature from the other ones and requires a distinct approach in order to be answered. For example, if we consider the last question: *What are people's feelings about Bush's decision?* The annotated answers are: *I am just disappointed that the US never supported it/ The whole world's perturbed with the greenhouse effect; emission gases destroying the earth and global warming causing terrible climate changes, except, of course President Bush./ After years of essentially rolling over against the Bush administration on*

Climate Change, the EU showed surprising spine. / A collection of reasons why we hate Bush.

Analyzing the answers we labeled in our corpus we can notice different problems. The first one is that we do not have a general opinion and, as a consequence, there is no correct or wrong answer. Each of them represents a different point of view about Bush.

The second one is that questions are written with words that are different from the ones of the answer, as for example synonyms or paraphrases, causing problems at the time of answer retrieval. Finally, we have to take into account that blog posts are written in a informal style and thus, we could be able to contemplate linguistic phenomena such as sayings or collocations that vary from one language to another. As we can deduce, opinion QA is a challenging task due to the fact that we add to the problems of general QA, the difficulties of the opinion mining research area, a complex field where linguistic phenomena, language features and subjectivity are involved.

DEFINING THE NECESSITIES OF OPINION QUESTION ANSWERING SYSTEMS – QUESTION ANALYSIS AND RETRIEVAL

Subsequently to defining a method to discriminate among factual and opinionated questions, we proceeded to studying the necessities of automatic systems that are able to correctly retrieve answers to opinionated questions (Balahur et al., 2009d). In order to carry out our evaluation, we employed the EmotiBlog corpus (Boldrini et al., 2009). It is a collection of blog entries in English, Spanish and Italian. However, for this research we used the first two languages. We annotated it using *EmotiBlog* and we also created a list of 20 questions for each language. Finally, we produced the *Gold Standard*, by labeling the corpus with the correct answers corresponding to the questions.

No	Type		Question
1	F	F	What international organization do people criticize for its policy on carbon emissions? *¿Cuál fue uno de los primeros países que se preocupó por el problema medioambiental?*
2	O	F	What motivates people's negative opinions on the Kyoto Protocol? *¿Cuál es el país con mayor responsabilidad de la contaminación mundial según la opinión pública?*
			What country do people praise for not signing the Kyoto

No	Type		Question
3	F	F	Protocol? *¿Quién piensa que la reducción de la contaminación se debería apoyar en los consejos de los científicos?*
4	F	F	What is the nation that brings most criticism to the Kyoto Protocol? *¿Qué administración actúa totalmente en contra de la lucha contra el cambio climático?*
5	O	F	What are the reasons for the success of the Kyoto Protocol? *¿Qué personaje importante está a favor de la colaboración del estado en la lucha contra el calentamiento global?*
6	O	F	What arguments do people bring for their criticism of media as far as the Kyoto Protocol is concerned? *¿A qué políticos americanos culpa la gente por la grave situación en la que se encuentra el planeta?*
7	O	F	Why do people criticize Richard Branson? *¿A quién reprocha la gente el fracaso del Protocolo de Kyoto?*
8	F	F	What president is criticized worldwide for his reaction to the Kyoto Protocol? *¿Quién acusa a China por provocar el mayor daño al medio ambiente?*
9	F	O	What American politician is thought to have developed bad environmental policies? *¿Cómo ven los expertos el futuro?*
10	F	O	What American politician has a positive opinion on the Kyoto protocol? *Cómo se considera el atentado del 11 de septiembre?*
11	O	O	What negative opinions do people have on Hilary Benn? *¿Cuál es la opinión sobre EEUU?*
12	O	O	Why do Americans praise Al Gore’s attitude towards the Kyoto protocol and other environmental issues? *¿De dónde viene la riqueza de EEUU?*
13	F	O	What country disregards the importance of the Kyoto Protocol? *¿Por qué la guerra es negativa?*
14	F	O	What country is thought to have rejected the Kyoto Protocol due to corruption? *¿Por qué Bush se retiró del Protocolo de Kyoto?*
			What alternative environmental friendly resources do people

No	Type		Question
15	F/O	O	suggest to use instead of gas en the future? *¿Cuál fue la posición de EEUU sobre el Protocolo de Kyoto?*
16	F/O	O	Is Arnold Schwarzenegger pro or against the reduction of CO2 emissions? *¿Qué piensa Bush sobre el cambio climático?*
17	F	O	What American politician supports the reduction of CO2 emissions? *¿Qué impresión da Bush?*
18	F/O	O	What improvements are proposed to the Kyoto Protocol? *¿Qué piensa China del calentamiento global?*
19	F/O	O	What is Bush accused of as far as political measures are concerned? *¿Cuál es la opinión de Rusia sobre el Protocolo de Kyoto?*
20	F/O	O	What initiative of an international body is thought to be a good continuation for the Kyoto Protocol? *¿Qué cree que es necesario hacer Yvo Boer?*

Table 5.12: List of question in English and Spanish

We created a set of factoid (F) and opinion (O) queries for English and for Spanish, presented in Table 5.12. Some of the questions could be defined between factoid and opinion (F/O) and the system can retrieve multiple answers after having selected, for example, the polarity of the sentences in the corpus.

The first step in our analysis was to evaluate and compare the generic QA system of the University of Alicante (Moreda et al., 2008) and the opinion QA system presented by Balahur et al. (2008), in which Named Entity Recognition with LingPipe[39] and FreeLing[40] was added, in order to boost the scores of answers containing NEs of the question Expected Answer Type (EAT).

The open domain QA system of the University of Alicante (Moreda et al., 2008) deals with factual questions as *location*, *person*, *organization*, *date-time* and *number* in English and also in Spanish. Its architecture comprises three modules:

- Question Analysis: the language object of the study is determined selecting the language for which more words are found. Further on, the question type is selected using a collection of regular expressions and the keywords of each question are obtained with morphological and dependencies analysis MINIPAR[41] for Spanish and Freeling[42] for English.

[39] http://alias-i.com/lingpipe/
[40] http://garraf.epsevg.upc.es/freeling/
[41] http://www.cs.ualberta.ca/~lindek/minipar.htm
[42] http://garraf.epsevg.upc.es/freeling/

- Information Retrieval: the system, originally, relied on the Internet search engines. However, in order to look for information among the Web Log collection, a keyword-based document retrieval method has been implemented to get relevant documents given the question keywords.
- Answer Extraction: the potential answers are selected using a NE recognizer for each retrieved document. LingPipe[43] and Freeling have been used for English and Spanish respectively. Furthermore, NE of the obtained question type and question keywords are marked up in the text. Once selected they are scored and ranked using answer-keywords distances approach. Finally, when all relevant documents have been explored, the system carries out an answer clustering process to group all answers that are equal or contained by others to the most scored one. Figure 5.3 presents the modules of QA system of the University of Alicante.

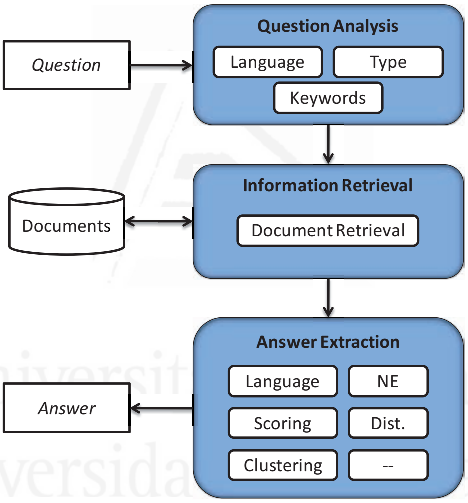


Figure 5.3: The architecture of the QA system of the University of Alicante

INITIAL METHOD FOR OPINION QUESTION ANSWERING

In order to test the performance of an opinion QA system, we used an approach similar to the one presented by Balahur et al. (2008).

Given an opinion question, we try to determine its polarity, the focus, its keywords (by eliminating stopwords) and the expected answer type (EAT) (while also marking the NE appearing in it). After having extracted this information from the question, we split blog texts into sentences and we also mark NEs. Finally, sentences in the blogs are sought which have the highest similarity score with the

[43] http://alias-i.com/lingpipe/

question keywords, whose polarity is the same as the determined question polarity and which contains a NE of the EAT. As the traditional QA system outputs 50 answers, we also take the 50 most similar sentences and extract the NEs they contain. In the future, when training examples will be available, we plan to set a threshold for similarity, thus not limiting the number of output answers, but setting a border to the similarity score (this is related to the observation made by Stoyanov et al. (2005) that opinion questions have a highly variable number of answers). The approach is depicted in Figure 5.4.

Further on, we present the details of our method. In order to extract the topic and determine the question polarity, we define question patterns. These patterns take into consideration the interrogation formula and extract the opinion words (nouns, verbs, adverbs, adjectives and their determiners). The opinion words are then classified in order to determine the polarity of the question, using the WordNet Affect emotion lists, the emotion triggers resource (Balahur and Montoyo, 2008), a list of four attitudes that we built, containing the verbs, nouns, adjectives and adverbs for the categories of criticism, support, admiration and rejection and a list of positive and negative opinion words taken from the system by Balahur and Montoyo (2008).

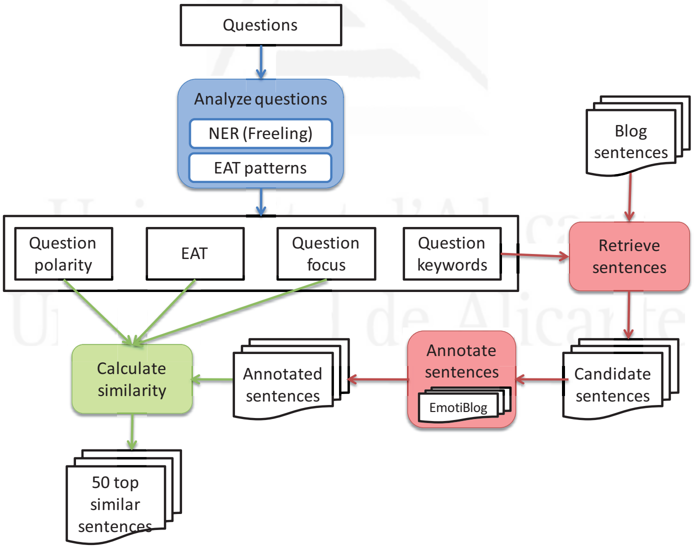


Figure 5.4: The opinion QA system

On the other hand, we pre-processed the blog texts in order to prepare the answer retrieval. Starting from the focus, keywords and topic of the question, we sought sentences in the blog collection (which was split into sentences and where Named Entity Recognition was performed using FreeLing) that could constitute possible answers to the questions, according to their similarity to the latter.

The similarity score was computed with Pedersen's Text Similarity Package[44] (using this software, both the words in the question, as well as the words in the blog sentences are also stemmed). The condition we subsequently set was that the polarity of the retrieved snipped be the same as the one of the question.

The polarity was computed using SVM on the trained model for the annotations in the EmotiBlog corpus, using as features the n-gram similarity (with n ranging from 1 to 4), as well as overall similarity to the annotated phrases, an approach similar to Balahur et al. (2009). Moreover, in the case of questions with EAT PERSON, ORGANIZATION or LOCATION, we required that a Named Entity of the appropriate type was present in the retrieved snippets and we boosted the score of the snippets fulfilling these conditions to the score of the highest ranking one. In case more than 50 snippets were retrieved, we only considered for evaluation the first 50 in the order of their similarity score, filtered by the polarity and NE presence criteria (which proved to be a good indicator of the snippet's importance (Balahur et al., 2008).

EVALUATION OF THE INITIAL APPROACH FOR OPINION QUESTION ANSWERING

Table 5.12 presents the results obtained for English and Table 5.13 for Spanish. We indicate the id of the question (Q), the question type (T) and the number of answer of the *Gold Standard* (A). We present the number of the retrieved questions by the traditional system (TQA) and by the opinion one (OQA). We take into account the first 1, 5, 10 and 50 answers.

Q	T	A	Number of answers found							
			@1		@5		@10		@ 50	
			TQA	OQA	TQA	OQA	TQA	OQA	TQA	OQA
1	F	5	0	0	0	2	0	3	4	4
2	O	5	0	0	0	1	0	1	0	3
3	F	2	1	1	2	1	2	1	2	1
4	F	10	1	1	2	1	6	2	10	4
5	O	11	0	0	0	0	0	0	0	0
6	O	2	0	0	0	0	0	1	0	2
7	O	5	0	0	0	0	0	1	0	3
8	F	5	1	0	3	1	3	1	5	1
9	F	5	0	1	0	2	0	2	1	3
10	F	2	1	0	1	0	1	1	2	1

[44] http://www.d.umn.edu/~tpederse/text-similarity.html

Q	T	A	Number of answers found							
			@1		@5		@10		@ 50	
			TQA	OQA	TQA	OQA	TQA	OQA	TQA	OQA
11	O	2	0	1	0	1	0	1	0	1
12	O	3	0	0	0	1	0	1	0	1
13	F	1	0	0	0	0	0	0	0	1
14	F	7	1	0	1	1	1	2	1	2
15	F/O	1	0	0	0	0	0	1	0	1
16	F/O	6	0	1	0	4	0	4	0	4
17	F	10	0	1	0	1	4	1	0	2
18	F/O	1	0	0	0	0	0	0	0	0
19	F/O	27	0	1	0	5	0	6	0	18
20	F/O	4	0	0	0	0	0	0	0	0

Table 5.12: Results of the opinion question answering method for English

Q	T	A	Number of answers found							
			@1		@5		@10		@ 50	
			TQA	OQA	TQA	OQA	TQA	OQA	TQA	OQA
1	F	9	1	0	0	1	1	1	1	3
2	F	13	0	1	2	3	0	6	11	7
3	F	2	0	1	0	2	0	2	2	2
4	F	1	0	0	0	0	0	0	1	0
5	F	3	0	0	0	0	0	0	1	0
6	F	2	0	0	0	1	0	1	2	1
7	F	4	0	0	0	0	1	0	4	0
8	F	1	0	0	0	0	0	0	1	0
9	O	5	0	1	0	2	0	2	0	4
10	O	2	0	0	0	0	0	0	0	0
11	O	5	0	0	0	1	0	2	0	3
12	O	2	0	0	0	1	0	1	0	1
13	O	8	0	1	0	2	0	2	0	4
14	O	25	0	1	0	2	0	4	0	8
15	O	36	0	1	0	2	0	6	0	15
16	O	23	0	0	0	0	0	0	0	0
17	O	50	0	1	0	5	0	6	0	10
18	O	10	0	1	0	1	0	2	0	2
19	O	4	0	1	0	1	0	1	0	1

Q	T	A	Number of answers found							
			@1		@5		@10		@ 50	
			TQA	OQA	TQA	OQA	TQA	OQA	TQA	OQA
20	O	4	0	1	0	1	0	1	0	1

Table 5.13: Results of the opinion question answering method for Spanish

DISCUSSION ON THE PRELIMINARY RESULTS

There are many problems involved when trying to perform mixed fact and opinion QA. The first can be the ambiguity of the questions e.g. *¿De dónde viene la riqueza de EEUU?*. The answer can be explicitly stated in one of the blog sentences, or a system might have to infer them from assumptions made by the bloggers and their comments. Moreover, most of the opinion questions have longer answers, not just a phrase snippet, but up to 2 or 3 sentences.

As we can observe in Table 5.12, the questions for which the TQA system performed better were the pure factual ones (1, 3, 4, 8, 10 and 14), although in some cases (question number 14) the OQA system retrieved more correct answers. At the same time, opinion queries, although revolving around NEs, were not answered by the traditional QA system, but were satisfactorily answered by the opinion QA system (2, 5, 6, 7, 11, 12). Questions 18 and 20 were not correctly answered by any of the two systems. We believe the reason is that question 18 was ambiguous as far as polarity of the opinions expressed in the answer snippets ("improvement" does not translate to either "positive" or "negative") and question 20 referred to the title of a project proposal that was not annotated by any of the tools used. Thus, the OQA system must be added a component for the identification of quotes and titles, as well as explore a wider range of polarity/opinion scales.

Furthermore, questions 15, 16, 18, 19 and 20 contain both factual as well as opinion aspects and the OQA system performed better than the TQA, although in some cases, answers were lost due to the artificial boosting of the queries containing NEs of the EAT (Expected Answer Type). Therefore, it is obvious that an extra method for answer ranking should be used, as Answer Validation techniques using Textual Entailment.

In Table 5.13, the OQA missed some of the answers due to erroneous sentence splitting, either separating text into two sentences where it was not the case or concatenating two consecutive sentences; thus missing out on one of two consecutively annotated answers. Examples are questions number 16 and 17, where many blog entries enumerated the different arguments in consecutive sentences. Another source of problems was the fact that we gave a high weight to the presence of the NE of the sought type within the retrieved snippet and in some cases the name was misspelled in the blog entries, whereas in other NER performed by FreeLing either attributed the wrong category to an entity, failed to annotate it or wrongfully annotated words as being NEs. Not of less importance is the question duality aspect in question 17. Bush is commented in more than 600 sentences; therefore, when polarity is not specified, it is difficult to correctly rank the answers. Finally, also the problems of temporal expressions and the co-reference need to be taken into account.

FINAL PROPOSAL FOR AN OPINION QUESTION ANSWERING SYSTEM

Summarizing our efforts until this point, subsequent to the research in classifying questions between factual and opinionated, we created a collection of both factual and opinion queries in Spanish and English. We labeled the Gold Standard of the answers in the corpora and subsequently we employed two QA systems, one open domain, one for opinion questions. Our main objective was to compare the performances of these two systems and analyze their errors, proposing solutions to creating an effective QA system for both factoid an opinionated queries. We saw that, even using specialized resources, the task of QA is still challenging. From our preliminary analysis, we could see that Opinion QA can benefit from snippet retrieval at a paragraph level, since in many cases the answers were not simple parts of sentences, but consisted in two or more consecutive sentences. On the other hand, we have seen cases in which each of three different consecutive sentences was a separate answer to a question.

Therefore, our subsequent efforts (Balahur et al., 2010a; Balahur et al., 2010c) concentrated on research to analyze the improvements that can be brought at the different stages of the OQA process: question treatment (identification of expected polarity – EPT, expected source – ES and expected target –ET-), opinion retrieval (at the level of one and three-sentences long snippets, using topic-related words or using paraphrases), opinion analysis (using topic detection and anaphora resolution). This research is motivated by the conclusions drawn by previous studies (Balahur et al., 2009d). Our purpose is to verify if the inclusion of new elements and methods - source and target detection (using semantic role labeling (SRL)), topic detection (using Latent Semantic Analysis) and joint topic-sentiment analysis (classification of the opinion expressed only in sentences related to the topic), followed by anaphora resolution (using a system whose performance is not optimal), affects the results of the system and how. Our contribution to this respect is the identification of the challenges related to OQA compared to traditional QA. We propose adding the appropriate methods, tools and resources to resolve the identified challenges. With the purpose of testing the effect of each tool, resource and technique, we will carry out a separate and a global evaluation. Until this point, although previous approaches opinion questions have longer answers than factual ones, the research done in OQA so far has only considered a sentence-level approach. This also includes our work (Balahur et al., 2009a; Balahur et al, 2009d). In the following experiments we will thus evaluate the impact of the retrieval at 1 and 3-sentence level and the retrieval based on similarity to query paraphrases enriched with topic-related words. We believe retrieving longer text could cause additional problems such as redundancy, co-reference and temporal expressions or the need to apply contextual information.

Starting from the research in previous works (Balahur et al., 2009a; Balahur et al., 2009d), our aim is to give a step forward and employ opinion specific methods focused on improving the performance of our OQA. We perform the retrieval at 1 sentence and 3 sentence-level and also determine the ES and the ET of the questions, which are fundamental to properly retrieve the correct answer. These two elements are selected employing SR. The expected answer type (EAT) is determined using Machine Learning (ML) using Support Vector Machines (SVM), by taking into account the interrogation formula, the subjectivity of the verb and the presence of polarity words in the target SR. In the case of expected opinionated answers, we also compute the EPT – by applying OM on the affirmative version of the question. These experiments are presented in more detail in the experiment section.

In order to carry out the present research for detecting and solving the complexities of opinion QA systems, we employed two blog posts corpora: EmotiBlog (Boldrini et al., 2009a) and the TAC 2008 Opinion Pilot test collection (part of the Blog06 corpus).

The TAC 2008 Opinion Pilot test collection is composed by documents with the answers to the opinion questions given on 25 targets. EmotiBlog is a collection of blog posts in English extracted from the Web. As a consequence, it represents a genuine example of this textual genre. It consists in a monothematic corpus about the Kyoto Protocol, annotated with the improved version of EmotiBlog (Boldrini et al., 2009b). It is well know that Opinion Mining (OM) is a very complex task due to the high variability of the language we study, thus our objective is to build an annotation model for an exhaustive detection of subjective speech, which can capture the most important linguistic phenomena, which give subjectivity to the text. Additional criteria employed when choosing the elements to be annotated were effectiveness and noise minimization. Thus, from the first version of the model, the elements not statistically relevant have been eliminated. The elements that compose the improved version of the annotation model are presented in Table 5.14.

Elements	Description
Obj. speech	Confidence, comment, source, target.
Subj. speech	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Adjectives/Adverbs	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Verbs/ Names	Confidence, comment, level, emotion, phenomenon, polarity, mode, source and target.
Anaphora	Confidence, comment, type, source and target.
Capital letter/ Punctuation	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Phenomenon	Confidence, comment, type, collocation, saying, slang, title, and rhetoric.
Reader/Author Interpretatipm (obj.)	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Emotions	Confidence, comment, accept, anger, anticipation, anxiety, appreciation, bad, bewilderment, comfort, compassion…

Table 5.14: EmotiBlog structure

The first distinction consists in objective and subjective speech. Subsequently, a finer-grained annotation is employed for each of the two types of data.

Objective sentences are annotated with source and target (when necessary also the level of confidence of the annotator and a comment).

The subjective elements can be annotated at a sentence level, but they also have to be labeled at a word level. *EmotiBlog* also contains annotations of anaphora at a cross-document level (to interpret the storyline of the posts) and the sentence type (simple sentence or title, but also saying or collocation).

Finally, the Reader and the Writer interpretation have to be marked in objective sentences. This element is employed to mark and interpret correctly an apparent objective discourse, whose aim is to implicitly express an opinion (e.g. “The camera broke in two days”). The first is useful to extract what is the interpretation of the reader (for example if the writer says *The result of their governing was an increase of 3.4% in the unemployment rate* instead of *The result of their governing was a disaster for the unemployment rate*) and the second to understand the background of the reader (i.e.. *These criminals are not able to govern* instead of saying *the x party is not able to govern*) from this sentence the reader can deduce the political ideas of the writer. The questions whose answers are annotated with EmotiBlog are the subset of opinion questions in English presented in (Balahur et al., 2009). The complete list of questions is shown in Table 5.15.

Question Number	Question
2	What motivates people’s negative opinions on the Kyoto Protocol?
5	What are the reasons for the success of the Kyoto Protocol?
6	What arguments do people bring for their criticism of media as far as the Kyoto Protocol is concerned?
7	Why do people criticize Richard Branson?
11	What negative opinions do people have on Hilary Benn?
12	Why do Americans praise Al Gore’s attitude towards the Kyoto protocol?
15	What alternative environmental friendly resources do people suggest to use instead of gas en the future?
16	Is Arnold Schwarzenegger pro or against the reduction of CO2 emissions?
18	What improvements are proposed to the Kyoto Protocol?
19	What is Bush accused of as far as political measures are concerned?
20	What initiative of an international body is thought to be a good continuation for the Kyoto Protocol?

Table 5.15: Questions over the EmotiBlog corpus

The main difference between the two corpora employed is that *Emotiblog* is monothematic, in fact only posts about the Kyoto Protocol compose it, while the TAC 2008 corpus contains documents on a multitude of subjects. Therefore, different techniques must be adjusted in order to treat each of them.

THE QUESTION ANALYSIS PHASE

In order to be able to extract the correct answer to opinion questions, different elements must be considered. As stated by Balahur et al. (2009), we need to determine both the expected answer type (EAT) of the question – as in the case of factoid ones - as well as new elements – such as expected polarity type (EPT). However, opinions are directional – i.e., they suppose the existence of a source and a target to which they are addressed.

Thus, we introduce two new elements in the question analysis – expected source (ES) and expected target (ET). These two elements are selected by applying SR and choosing the source as the agent in the sentence and the direct object (patient) as the target of the opinion. The expected answer type (EAT) (e.g. opinion or other) is determined using Machine Learning (ML) using Support Vector Machine (SVM), by taking into account the interrogation formula, the subjectivity of the verb and the presence of polarity words in the target SR. In the case of expected opinionated answers, we also compute the expected polarity type (EPT) – by applying OM on the affirmative version of the question. An example of such a transformation is: given the question *“What are the reasons for the success of the Kyoto Protocol?*,” the affirmative version of the question is *“The reasons for the success of the Kyoto Protocol are X”.*

In the answer retrieval stage, we employ four strategies:

1. Using the JIRS (JAVA Information Retrieval System) IR engine (Gómez et al., 2007) to find relevant snippets. JIRS retrieves passages (of the desired length), based on searching the question structures (n-grams) instead of the keywords, and comparing them.
2. Using the “Yahoo” search engine to retrieve the first 20 documents that are most related to the query. Subsequently, we apply LSA on the retrieved documents and extract the words that are most related to the topic. Finally, we expand the query using words that are very similar to the topic and retrieve snippets that contain at least one of them and the ET.
3. Generating equivalent expressions for the query, using the DIRT paraphrase collection (Lin and Pantel, 2001) and retrieving candidate snippets of length 1 and 3 (length refers to the number of sentences retrieved) that are similar to each of the new generated queries and contain the ET. Similarity is computed using the cosine measure. Examples of alternative queries for *“People like George Clooney”* are *“People adore George Clooney”, “People enjoy George Clooney”, “People prefer George Clooney”.*
4. Enriching the equivalent expressions for the query in 3. with the topic-related words discovered in 2. using LSA.

In order to determine the correct answers from the collection of retrieved snippets, we must filter only the candidates that have the same polarity as the question EPT. For polarity detection, we use a combined system employing SVM ML on unigram and bigram features trained on the NTCIR MOAT 7 data and an unsupervised lexicon-based system. In order to compute the features for each of the unigrams and bigrams, we compute the tf-idf scores.

The unsupervised system uses the Opinion Finder lexicon to filter out subjective sentences – that contain more than two subjective words or a subjective word and a valence shifter (obtained from the General Inquirer resource). Subsequently, it accounts for the presence of opinionated words from four different lexicons – Micro WordNet (Cerini et al., 2007), WNAffect (Strapparava and Valitutti, 2004), Emotion Triggers (Balahur and Montoyo, 2008) and General Inquirer (Stone et al., 1966). For the joint topic-polarity analysis, we first employ LSA to determine the words that are strongly associated to the topic. Consequently, we compute the polarity of the sentences that contain at least one topic word and the question target.

Finally, answers are filtered using the Semrol system for SR labeling proposed by Moreda (2008). Subsequently, we filter all snippets that have the required target and source as agent or patient. Semrol receives as input plain text with information about grammar, syntax, word senses, Named Entities and constituents of each verb. The system output is the given text, in which the semantic roles information of each constituent is marked. Ambiguity is resolved depending on the machine algorithm employed, which in this case is TIMBL[45].

5.2.4. EVALUATION OF THE PROPOSED OPINION QUESTION ANSWERING FRAMEWORK

We evaluate our approaches on both the EmotiBlog question collection, as well as the TAC 2008 Opinion Pilot test set. We compare them against the performance of the system proposed by Balahur et al. (Balahur et al., 2009d) and the best (Copek et al., 2008) and lowest-scoring (Varma et al., 2008) systems as far as F-measure is concerned in the TAC 2008 task. For both the TAC 2008 and EmotiBlog sets of questions, we employ the SR system in SA and determine the ES, ET and EPT. Subsequently, for each of the two corpora, we retrieve 1-phrase and 3-phrase snippets. The retrieval of the of the EmotiBlog candidate snippets is done using query expansion with LSA and filtering according to the ET. Further on, we apply sentiment analysis (SA) approach and select only the snippets whose polarity is the same as the determined question EPT. The results are presented in Table 5.16.

Q No.	**No.** A	**Baseline** (Balahur et al., 2009d)				**1 phrase + ET+SA**				**3 phrases +ET+SA**			
		@ 1	**@ 5**	**@ 10**	**@ 50**	**@ 1**	**@ 5**	**@ 10**	**@ 50**	**@ 1**	**@ 5**	**@ 10**	**@ 20**
2	5	0	2	3	4	1	2	3	4	1	2	3	4
5	11	0	0	0	0	0	2	2	2	1	2	3	4
6	2	0	0	1	2	1	1	2	2	0	1	2	2
7	5	0	0	1	3	1	1	1	3	0	2	2	4
11	2	1	1	1	1	0	0	0	0	0	0	0	1
12	3	0	1	1	1	0	1	2	3	0	0	1	2
15	1	0	0	1	1	0	0	1	1	1	1	1	1

[45] http://ilk.uvt.nl/downloads/pub/papers/Timbl_6.2_Manual.pdf and http://ilk.uvt.nl/timbl/

Q No.	**No.** A	**Baseline** (Balahur et al., 2009d)				**1 phrase + ET+SA**				**3 phrases +ET+SA**			
16	6	1	4	4	4	0	1	1	2	1	2	2	6
18	1	0	0	0	0	0	0	0	0	0	0	0	0
19	27	1	5	6	18	0	1	1	2	0	1	1	1
20	4	0	0	0	0	0	0	1	1	0	0	1	2

Table 5.16: Results for questions over EmotiBlog

The retrieval of the TAC 2008 1-phrase and 3-phrase candidate snippets was done using JIRS. Subsequently, we performed different evaluations, in order to assess the impact of using different resources and tools. Since the TAC 2008 had a limit of the output of 7000 characters, in order to compute a comparable F-measure, at the end of each processing chain, we only considered the snippets for the 1-phrase retrieval and for the 3-phases one until this limit was reached.

1. In the first evaluation, we only apply the sentiment analysis tool and select the snippets that have the same polarity as the question EPT and the ET is found in the snippet. (i.e. *What motivates peoples negative opinions on the Kyoto Protocol? The Kyoto Protocol becomes deterrence to economic development and international cooperation/ Secondly, in terms of administrative aspect, the Kyoto Protocol is difficult to implement.* - same EPT and ET)
   We also detected cases of same polarity but no ET, e.g. *These attempts mean annual expenditures of $700 million in tax credits in order to endorse technologies, $3 billion in developing research and $200 million in settling technology into developing countries* –EPT negative but not same ET.
2. In the second evaluation, we add the result of the LSA process to filter out the snippets from 1., containing the words related to the topic starting from the retrieval performed by Yahoo, which extracts the first 20 documents about the topic.
3. In the third evaluation, we filter the results in 2 by applying the *Semrol* system and setting the condition that the ET and ES are the agent or the patient of the snippet.
4. In the fourth evaluation setting, we replaced the set of snippets retrieved using JIRS with the ones obtained by generating alternative queries using paraphrases. We subsequently filtered these results based on their polarity (so that it corresponds to the EPT) and on the condition that the source and target of the opinion (identified through SRL using Semrol) correspond to the ES and ET.
5. In the fourth evaluation setting, we replaced the set of snippets retrieved using JIRS with the ones obtained by generating alternative queries using

paraphrases, enriched with the topic words determined using LSA. We subsequently filtered these results based on their polarity (so that it corresponds to the EPT) and on the condition that the source and target of the opinion (identified through SRL using Semrol) correspond to the ES and ET.

System	F-measure
Best TAC	0.534
Worst TAC	0.101
JIRS + SA+ET (1 phrase)	0.377
JIRS + SA+ET (3 phrases)	0.431
JIRS + SA+ET+LSA (1 phrase)	0.489
JIRS + SA+ET+LSA (3 phrases)	0.505
JIRS + SA+ET+LSA+SR (1 phrase)	0. 533
JIRS + SA+ET+LSA+SR (3 phrases)	0.571
PAR+SA+ET+SR(1 phrase)	0.345
PAR+SA+ET+SR(2 phrase)	0.386
PAR_LSA+SA+ET+SR (1 phrase)	0.453
PAR_LSA+SA+ET+SR (3 phrases)	0.434

Table 5.17: Results for the TAC 2008 question set

From the results obtained, we can draw the following conclusions. Firstly, the hypothesis that OQA requires the retrieval of longer snippets was confirmed by the improved results, both in the case of *EmotiBlog*, as well as the TAC 2008 corpus. Secondly, opinion questions require the joint topic-sentiment analysis; as we can see from the results, the use of topic-related words in the computing of the affect influences the results in a positive manner and joint topic-sentiment analysis is especially useful for the cases of questions asked on a monothematic corpus. Thirdly, another conclusion that we can draw is that target and source detection is a relevant step at the time of answer filtering, not only helping in the more accurate retrieval of answers, but also at placing at the top of the retrieval the relevant results. Nonetheless, as we can see from the relatively low improvement in the results, much remains to be done in order to appropriately tackle OQA. As seen in the results, there are still questions for which no answer is found (e.g. 18). This is due to the fact that its treatment requires the use of inference techniques that are presently unavailable (i.e. define terms such as “improvement”).

The results obtained when using all the components, for the 3-sentence long snippets significantly improve the results obtained by the best system participating in the TAC 2008 Opinion Pilot competition (determined using a paired t-test for statistical significance, with confidence level 5%). Finally, from the analysis of the errors, we could see that even though some tools are in theory useful and should produce higher improvements – such as SR – their performance in reality does not produce drastically higher results. The idea to use paraphrases for query expansion also proved to decrease the system performance. From preliminary results obtained using JavaRap[46] for co- reference resolution, we also noticed that the performance of the OQA lowered, although theoretically it should have improved.

With the objective of improving the task of QA in the context of opinion data, we presented and evaluated different methods and techniques. From the evaluations performed using different NLP resources and tools, we concluded that joint topic-sentiment analysis, as well as the target and source identification, are crucial for the correct performance of this task. We have also demonstrated that by retrieving longer answers, the results have improved. We thus showed that opinion QA requires the development of appropriate strategies at the different stages of the task (recognition of subjective question, detection of subjective content of the question, source, and target and retrieving of the required data). Due to the high level of complexity of the subjective language, further improvements could be obtained by testing higher-performing tools for co-reference resolution, other (opinion) paraphrases collections and paraphrasing methods and the employment of external knowledge sources that refine the semantics of queries. Another important issue to be solved is the resolution of temporal expression. This issue is tackled in the next section.

5.2.5. OPINION QUESTION ANSWERING WITH TEMPORAL RESTRICTIONS – PARTICIPATION IN THE NTCIR 8 MOAT

Having analyzed the needs of an opinion question answering system and proposed adequate solutions for tackling this task, we subsequently aimed at evaluating our approach in an open competition. This external evaluation of our approaches was done in the NTCIR 8 MOAT (Multilingual Opinion Analysis Task).

In this competition, the participants were provided with twenty topics. For each of the topics, a question was given, together with a short and concise query corresponding to the question, the expected polarity of the answer and the period of time required. For each of the topics, the participants were given a set of documents, that were split into sentences (for the opinionated and relevance judgments) and into opinion units (for the polarity, opinion target and source tasks). 5 different subtasks were defined both in a monolingual, as well as cross-language setting: judging sentence opinionatedness, relevance, determining the polarity of opinionated sentence, as well as the source and target of the opinions identified. The monolingual subtasks were defined for 4 languages: English, Traditional Chinese, Simplified Chinese and Japanese. In the cross-lingual task, participants

[46] http://wing.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.htm

could retrieve answers to the set of questions in English in any of the other 3 languages.

In order to evaluate our approaches to opinion question answering, as well as evaluate the inclusion of filtering techniques based on topic relevance and temporal restrictions, we participated in the first 3 subtasks in an English, monolingual setting, as well as in the cross-lingual challenge, retrieving the answers to the question set in English in the document set in Traditional Chinese. For the participation in this competition, we named our opinion question answering system OpAL.

For the English monolingual subtasks, we submitted three runs of the OpAL system, for the opinionated, relevance and polarity judgment tasks.

A) TACKLING THE ENGLISH MONOLINGUAL SUBTASKS AT NTCIR 8 MOAT

Judging sentence opinionatedness

The “opinionated” subtask required systems to assign the values YES or NO (Y/N) to each of the sentences in the document collection provided. This value is given depending on whether the sentence contains an opinion (Y) or it does not (N).

In order to judge the opinionatedness of the sentence, we employed two different approaches (the first one corresponding to system run number 1 and the second to system runs 2 and 3).

Both approaches are rule-based, but they differ in the resources employed. We considered as opinionated sentences the ones that contain at least two opinion words or one opinion word preceded by a modifier. For the first approach, the opinion words were taken from the General Inquirer, Micro WordNet Opinion and Opinion Finder lexicon and in the second approach we only used the first two resources.

Determining sentence relevance

In the sentence relevance judgment task, the systems had to output, for each sentence in the given collection documents per topic, an assessment on whether or not the sentence is relevant for the given question. For the sentence relevance judgement task stage, we employ three strategies (corresponding to the system runs 1, 2 and 3, respectively):

1. Using the JIRS (JAVA Information Retrieval System) IR engine (Gómez et al., 2007) to find relevant snippets. JIRS retrieves passages (of the desired length), based on searching the question structures (n-grams) instead of the keywords, and comparing them.

2. Using faceted search in Wikipedia and performing Latent Semantic Analysis (LSA) to find the words that are most related to the topic. The idea behind this approach is to find the concepts that are contained in the query descriptions of the topics. In order to perform this task, we match the query words, starting from the first, to a category in Wikipedia. Subsequently we match each group of two consecutive words to the same categories, then groups of 3, 4, etc. until the highest match is found. The concepts determined through this process are considered as the topic components. For each of these topic components, we determine the most related words, applying LSA is to the first 20 documents that are retrieved using the Yahoo search engine, given the query. For LSA, we employ the Infomap NLP[47] software.

   Finally, we expand query using words that are very similar to the topic (retrieved through the LSA process) and retrieve snippets that contain at least two such words.
3. The third approach consists in judging, apart from the topic relevance characteristic, the temporal appropriateness of the given sentences. In order to perform this check, we employ TERSEO (Saquete et al., 2006). We then filter the sentences obtained in the second approach depending on whether or not the document in which they appear have a date matching the required time interval or the sentence with the resolved temporal expressions contains a reference to the required time interval.

Polarity and topic-polarity classification for judging sentence answerness

The polarity judgment task required the system to assign a value of POS, NEG or NEU (positive, negative or neutral) to each of the sentences in the documents provided. In order to determine the polarity of the sentences, we passed each sentence through an opinion mining system employing SVM machine learning over the NTCIR 7 MOAT corpus, the MPQA corpus and EmotiBlog. Each sentence is preprocessed using Minipar[48]. For the system training, the following features were considered, for each sentence word:

- Part of speech (POS);
- Opinionatedness/intensity - if the word is annotated as opinion word, its polarity, i.e. 1 and -1 if the word is positive or negative, respectively and 0 if it is not an opinion word, its intensity (1.2 or 3) and 0 if it is not a subjective word, its emotion (if it has, none otherwise);

[47] http://infomap-nlp.sourceforge.net/
[48] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm

- Syntactic relatedness with other opinion word – if it is directly dependent of an opinion word or modifier (0 or 1), plus the polarity/intensity and emotion of this word (0 for all the components otherwise).

The difference between the submitted runs consisted in the lexicons used to determine whether a word was opinionated or not. For the first run, we employed the General Inquirer, MicroWordNet and the Opinion Finder opinion resources. For the second one, we employed, aside from these three sources, the "emotion trigger" resource (Balahur and Montoyo, 2008).

B) TACKLING THE ENGLISH-CHINESE CROSS-LINGUAL SUBTASK AT NTCIR 8 MOAT

In the Cross-lingual setting, the task of the participating systems was to output, for each of the twenty topics and their corresponding questions (in a language), the list of sentences containing answers (in another language). For this task, we submitted three runs of the OpAL system, all of them for the English- Traditional Chinese cross-lingual setting (i.e. the topics and questions are given in English; the output of the system contains the sentences in set of documents in Traditional Chinese which contain an answer to the given topics).

In the following part, we explain the approaches we followed for each of the system runs. Given that we had no previous experience with processing Chinese text, the approaches taken were quite simple.

The first step we performed was to tokenize the Chinese texts using LingPipe[49]. Further on, we applied a technique known as "triangulation" to obtain opinion and subjectivity resources for Chinese. The idea behind this approach is to obtain resources for different languages, starting from correct parallel resources in 2 initial languages. The process is exemplified in Figure 5.5 for obtaining resources in Chinese, starting with resources in English and Spanish.

As mentioned before, this technique requires the existence of two correct parallel resources in two different languages to obtain correct resources for a third language. We have previously translated and cleaned the General Inquirer[50], MicroWordNet and Opinion Finder lexicons for Spanish. The "emotion triggers" resource is available both for English, as well as for Spanish. In order to obtain these resources for Traditional Chinese, we use the Google translator.

[49] http://alias-i.com/lingpipe/

[50] http://www.wjh.harvard.edu/~inquirer/

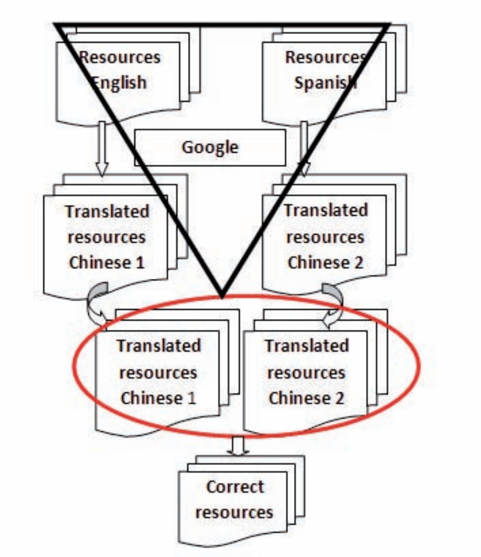

Resources English
Resources Spanish
Google
Translated resources Chinese 1
Translated resources Chinese 2
Translated resources Chinese 1
Translated resources Chinese 2
Correct resources

Figure 5.5: Obtaining new resources in Chinese through triangulation

We translate both the English, as well as the Spanish resources, into Traditional Chinese. Subsequently, we performed the intersection of the obtained translations – that is, the corresponding words that have been translated in the same manner – both from English as well as from Spanish. We removed words that we translated differently from English and Spanish. The intersection words were considered as “clean” (correct) translations. We mapped each of these resources to four classes, depending on the score they are assigned in the original resource – of “high positive”, “positive”, “high negative” and “negative” and we give each word a corresponding value (4, 1, -4 and -1), respectively.

On the other hand, we translated the topic words determined in English using LSA. For each of the sentence, we compute a score, given by the sum of the values of the opinion words that are matched in it.

In order for a sentence to be considered as answer to the given question, we set the additional conditions that it contains at least one topic word and that the polarity determined corresponds to the required polarity, as given in the topic description.

The three runs differ in the resources that were employed to calculate the sentiment score: in the first run, we employed the General Inquirer and MicroWordNet resources; in the second run we added the “emotion trigger resource” and the third run used only the Opinion Finder lexicon.

The following tables present the results of the system runs for the three subtasks in English in which we took part and the cross-lingual English - Traditional Chinese task.

System RunID	P	R	F
OpAL 1	17.99	45.16	25.73
OpAL 2	19.44	44	26.97
OpAL 3	19.44	44	26.97

Table 5.18: Results of system runs for opinionatedness

System RunID	P	R	F
OpAL 1	82.05	47.83	60.43
OpAL 2	82.61	5.16	9.71
OpAL 3	76.32	3.94	7.49

Table 5.19: Results of system runs for relevance

System RunID	P	R	F
OpAL 1	38.13	12.82	19.19
OpAL 2	50.93	12.26	19.76

Table 5.20: Results of system runs for polarity

System RunID	P	R	F
OpAL 1	3.54	56.23	6.34
OpAL 2	3.35	42.75	5.78
OpAL 3	3.42	72.13	6.32

Table 5.21: Results of system runs for the cross-lingual task – agreed measures, Traditional Chinese

System RunID	P	R	F
OpAL 1	14.62	60.47	21.36
OpAL 2	14.64	49.73	19.57
OpAL 3	15.02	77.68	23.55

Table 5.22: Results of system runs for the cross-lingual task – non-agreed measures, Traditional Chinese

DISCUSSION AND CONCLUSIONS

From the results obtained, on the one hand, we can see that although the extensive filtering according to the topic and the temporal restrictions increases the system precision, we obtain a dramatic drop in the recall. On the other hand, the use of simpler methods in the cross-lingual task yielded better results, the OpAL cross-lingual run 3 obtaining the highest F score for the non-agreed measures and ranking second according to the agreed measures.

From the error analysis performed, we realized that, on the one hand, the LSA-based method to determine topic-related words is not enough to perform this task. The terms obtained by employing this method are correct and useful, but they should be expanded using language models, to better account for the language variability.

Finally, we have seen that systems performing finer tasks, such as temporal expression resolution, are not mature enough to be employed in such tasks. This was confirmed by in-house experiments using anaphora resolution tools such as JavaRAP[51], whose use also led to lower performances of the system and dramatic loss in recall.

5.2.6. CONCLUSIONS

In this section, our research was focused on solving a recent problem born with the massive usage of the Web 2.0: the exponential growth of the opinionated data that need to be efficiently managed for a wide range of practical applications.

We identified and explored the challenges raised by OQA, as opposed to the traditional QA. Moreover, we studied the performance of new sentiment-topic detection methods and analyzed the improvements that can be brought at the different stages of the OQA process and analyzed the contribution of discourse analysis, employing techniques such as co-reference resolution and temporality detection. We also experimented new retrieval techniques such as faceted search using Wikipedia with LSA, which demonstrate to improve the performance of the task.

From the results obtained, we can draw the following conclusions. The first one is that on the one hand, the extensive filtering according to the topic and the temporal restrictions increases the system precision but it produces a dramatic drop in the recall. As a consequence, the use of simpler methods in the cross-lingual task would be more appropriate in this context. The OpAL cross-lingual run 3 obtaining the highest F score for the non-agreed measures and ranking second according to the agreed measures. On the other hand, we can deduce that LSA-based method to determine topic-related words is not enough to perform this task. The terms obtained by employing this method are correct and useful; however, as future work our purpose is to use language models, to better account for the language variability. Finally, we understand that co-reference or temporal resolution systems

[51] http://aye.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html

do not improve the performance of OQA, and as a consequence there is a need to study the performance of other co-reference and temporal resolution systems in order to check if the technique is not enough mature or if other systems can bring added value to this task.

5.3. OPINION-ORIENTED SUMMARIZATION FOR MASS OPINION ESTIMATION

5.3.1. INTRODUCTION AND BACKGROUND

The research described until this point has been motivated by the participation in the TAC 2008 Opinion Pilot competition, where systems were supposed to create summaries from the answers to opinion questions.

Nevertheless, as we outlined in the previous chapter, the need to summarize opinions is not specific only to answers to opinion questions, but a requirement to any system that aims at processing opinionated content, in order to support a user’s need for information extracted from such type of data. In order to be of real help to users, the automatic systems that are able to detect opinions must be enhanced with a summarization component, in order to deal with the remaining issue after opinion data has been mined and classified: large volume and high redundancy. In blogs, for example, an opinion mining system aiming at analyzing threads (the sequence of texts containing the post on a subject and the subsequent comments on it made by different “bloggers"), should analyze its content as far as opinion is concerned. However, the result obtained after this processing would still contain a large quantity of redundant information. Therefore, in order for the application to be truly useful to user, another component is needed that subsequently summarizes the classes of opinions expressed – e.g. arguments pro and against the topic.

In the first part of this chapter, we defined, employed and evaluated different methods for the summarization of answers to opinion questions that were retrieved from blogs. However, there are several issues related to evaluating the methods we proposed in this scenario:

- The evaluation of the summarization process is dependent on the retrieval and opinion mining stages, so the opinion summarization process cannot be directly evaluated;
- The snippets that should be contained in the final summary do not have a polarity associated to them, so it is impossible to evaluate the correctness of the opinion mining component;

Therefore, the aim of this subsequent research is to study the manner in which opinion can be summarized, so that the obtained summary can be used in real-life applications e.g. marketing, decision-making. We discuss the aspects involved in this task and the challenges it implies, in comparison to traditional text summarization, demonstrating how and why it is different from content-based summarization. We propose the labeling of a corpus for this task, as well as three different approaches to perform opinion summarization and test our hypotheses.

Additionally, we compare and evaluate the results of employing opinion mining versus summarization as a first step in opinion summarization. Subsequently, we propose an adequate method to tackle the summarization of opinions, in order to create high-performance systems for real-world applications.

5.3.2. INITIAL EXPERIMENTS

CORPORA

The first corpus we employed in our experiments is a collection of 51 blog threads extracted from the Web (Balahur et al., 2009a), which was labeled with EmotiBlog. The structure of the data is presented in Table 5.23. The elements from EmotiBlog used for the labeling of the data are presented in Table 5.24.

	Number of Posts	**Number of words per news item**	**Number of words per post**	**Total number of words**
Total	1829	72.995	226.573	299.568
Average	33.87	1351.75	4195.79	5547.55

Table 5.23: Structure of the corpus annotated for blog thread summarization

Element	**Attribute**
Polarity	Positive, negative
Level	Low, medium, high
Source	name
Target	name

Table 5.24: Elements from the EmotiBlog scheme used in for the annotation of blog threads

The blog threads are written in English and have the same structure: the authors create an initial post containing a piece of news and possibly their opinion on it and subsequently, bloggers reply, expressing their opinions about the topic (thus forming a discussion thread). The blog corpus annotation contains the URL from which the thread was extracted, the initial annotated piece of news and the labeled user comments. The topics contained in the corpus are very diverse: economy, science and technology, cooking, society and sports. The data is annotated at a document level, with the overall polarity and topic, and at sentence level, discriminating between objective and subjective sentences. Subsequently, subjective sentences are annotated with the polarity of the sentiment expressed and the intensity of the opinion expressed (low, medium or high). Finally, the source of the discourse and the target of the sentence are specified. Figure 5.6 contains an example of annotation. We would like to stress upon the fact that we indicate more than one topic. We decided to contemplate cases of multiple topics only if they are relevant in the blog. In this case, the main topic is the economic situation, while the secondary ones are the government and banks.

```
<topic>economic situation</topic>
<topic2>government</topic2>
<topic3>banks</topic3>
<news> Saturday, May 9, 2009 My aim in this blog has largely been to
give my best and most rational perspective on the reality of the
economic situation. I have tried (and I hope) mostly succeeded in
avoiding emotive and partisan viewpoints, and have tried as far as
possible to see the actions of politicians as misguided. Of late,
that perspective has been slipping, for the UK, the US and also for
Europe. </news>
<phenomenon gate:gateId="1" target="economic crisis" degree1="medium"
category="phrase" source="Cynicus Economicus" polarity1="negative" >I
think that the key turning point was the Darling budget, in which the
forecasts were so optimistic as to be beyond any rational
belief</phenomenon>…
```

```
<topic>economic situation</topic>
<topic2>government</topic2>
<topic3>banks</topic3>
<news> Saturday, May 9, 2009 My aim in this blog has largely been to
give my best and most rational perspective on the reality of the
economic situation. I have tried (and I hope) mostly succeeded in
avoiding emotive and partisan viewpoints, and have tried as far as
possible to see the actions of politicians as misguided. Of late,
that perspective has been slipping, for the UK, the US and also for
Europe. </news>
<phenomenon gate:gateId="1" target="economic crisis" degree1="medium"
category="phrase" source="Cynicus Economicus" polarity1="negative" >I
think that the key turning point was the Darling budget, in which the
forecasts were so optimistic as to be beyond any rational
belief</phenomenon>…
```

Figure 5.6:Example of annotation of blog thread using EmotiBlog

The second dataset used in our experiments is the collection of 85 bank reviews with fine-grained classification used by Saggion and Funk (2010). The aim in evaluating our approach on this dataset is to verify whether, in the context of opinion summarization, the first step should be opinion mining or summarization and how this choice in influences the final output of the system.

The main objective of our experiments is to design a system that is able to produce opinion summaries, in two different types of texts: a) blog threads, in which case we aim at producing summaries of the positive and negative arguments given on the thread topic; and b) reviews, in the context of which we assess the best manner to use opinion summarization in order to determine the overall polarity of the sentiment expressed. In our first opinion summarization experiments, we adopt a standard approach by employing in tandem a sentiment classification system and a text summarizer. The output of the former is used to divide the sentences in the blog threads into three groups: sentences containing positive sentiment, sentences containing negative sentiment and neutral or objective sentences. Subsequently, the positive and the negative sentences are passed on to the summarizer separately to produce one summary for the positive posts and another one for the negative ones. Next, we present the sentiment analysis system followed by a description of the summarization system, both of which serve as a foundation for subsequent sections. The ideas and results presented in this section were initially put forward in Balahur et al. (Balahur et al., 2009g).

The Sentiment Analysis System

The first step we took in our approach was to determine the opinionated sentences, assign each of them a polarity (positive or negative) and a numerical value corresponding to the polarity strength (the higher the negative score, the more negative the sentence and vice versa). Given that we are faced with the task of classifying opinion in a general context, we employed the simple, yet efficient approach, presented in Balahur et al. (2009).

In the following experiments, we used WordNet Affect (Strapparava and Valitutti, 2004), SentiWordNet (Esuli and Sebastiani, 2005), MicroWNOp (Cerini et al., 2007). Each of the resources we employed were mapped to four categories, which were given different scores: positive (1), negative (-1), high positive (4) and high negative (4).

As we have shown (Balahur et al., 2009f), these values per formed better than the usual assignment of only positive (1) and negative (-1) values. First, the score of each of the blog posts was computed as the sum of the values of the words that were identified; a positive score leads to the classification of the post as positive, whereas a negative score leads to the system classifying the post as negative.

Subsequently, we performed sentence splitting using Lingpipe and classified the sentences we thus obtained according to their polarity, by adding the individual scores of the affective words identified.

The Summarization System

The summarization system we employed in this preliminary analysis, presented by Balahur et al. (Balahur et al., 2009g) is the one described by Steinberger and Ježek (2008). It was originally proposed by Gong and Liu (2002) and later improved by Steinberger and Ježek (2004). The main idea of the approach relies on the use of Latent Semantic Analysis (LSA) to detect the topic-relevant sentences. This approach first builds a term-by-sentence matrix from the source, then applies Singular Value Decomposition (SVD) and finally uses the resulting matrices to identify and extract the most salient sentences. SVD extends the latent (orthogonal) dimensions, which in simple terms correspond to the different topics discussed in the source.

EVALUATION OF THE INITIAL OPINION SUMMARIZATION PROPOSAL

In order to test the performance of this first approach, we first apply the sentiment analysis system proposed by Balahur et al. (Balahur et al., 2009f) and subsequently the text summarizer described by Steinberger and Ježek (2008). We analyze and discuss the performance of the sentiment recognition, followed by the overall summarization performance. Performance results of the sentiment analysis are shown in Table 5.25. We have also analyzed in depth the results of 14 blog threads in order to assess the quality of the output produced by the opinion mining system, independently of the topic relevance of the sentences classified. As only sentences that were relevant to the topic in question were labeled in the Gold Standard, we also assessed the sentiment of the sentences that were not annotated (see Table 5.26).

System	Precision	Recall	F1
Sent$_{pos}$	0.53	0.89	0.67
Sent$_{neg}$	0.67	0.22	0.33

Table 5.25: Results of the evaluation of the sentiment analysis system

As we can observe from the results presented in Table 5.25, the system we employed had an overall relatively low precision and a high recall (measured as average of the results obtained for the positive and negative sets of summaries), meaning that the sentences that were classified as positive or negative by the system were either erroneously classified (as positive, when they were in fact negative, or vice-versa) or they were annotated in the Gold Standard as being objective. However, the high recall suggests that the system, although simple, is capable of distinguishing subjective sentences from objective ones. We further on analyzed the performance of the sentiment analysis system independently of the topic relevance. The results are summarized in Table 5.26.

Thread No.	Number of sentences with positive sentiment				Number of sentences with negative sentiment				Asserted Pos	Asserted Neg
	Total	O-P	O-N	O-O	Total	O-P	O-N	O-O		
1	5	0	1	2	4	0	2	0	2	2
2	5	0	0	1	5	1	2	1	1	1
3	4	1	2	1	4	0	2	0	0	2
4	4	2	1	1	4	0	2	1	0	1
6	5	1	4	0	5	0	4	1	0	1
8	5	4	0	1	6	0	3	2	0	1
11	4	3	0	1	6	3	3	0	0	0
12	4	0	3	1	6	0	3	1	0	2
14	5	1	3	1	5	0	4	0	0	1
15	5	2	2	1	5	0	3	2	0	0
16	4	2	2	0	5	1	2	0	0	2
18	5	0	3	2	4	1	0	3	0	0
30	7	3	3	1	4	0	2	0	0	2
31	5	2	3	0	5	1	3	0	0	1

Table 5.26: Results of the evaluation of the sentiment analysis classification performance on 14 topics, independently of the topic relevance

As we can observe from the results presented in Table 5.26, the opinion mining system performed well as far as polarity classification was concerned. Thus, although relatively few of the sentences in the summary are also present in the Gold Standard as far as polarity and sentence importance are concerned, the sentences were classified well, especially for the negative class (see correlation between Negative and O-N and Positive and O-P in Table 5.26). The degree of importance of each 'latent' topic is given by the singular values and the optimal number of latent topics (i.e., dimensions) r can be re-tuned on training data. Thus, some improvement can be achieved over these results by adding a topic detection component.

It can also be noticed that the system has a tendency to over classify sentences as being negative, a fault which we attribute to the fact that in our approach we do not contemplate negations and the fact that the resources used contain, in their original form, word senses, and we do not perform any word sense disambiguation. Additionally, most of the resources used for sentiment detection have a large number of negative terms and a significantly lower number of positive ones.

Performance results of the summarizer are shown in Table 5.27. We used the standard ROUGE evaluation (20) which has also been used for the Text Analysis

Conference (TAC) series of competitions. We include the usual ROUGE metrics: R1 is the maximum number of co-occurring unigrams, R2 is the maximum number of co-occurring bigrams, RSU4, is the skip bigram measure with the addition of unigrams as counting unit, and finally, RL is the longest common subsequence measure. In all cases, we present the average F1 score for the given metric and the 95% confidence intervals within parenthesis. There are three rows in Table 5.27: the first one (Sent+Summ$_{neg}$) corresponds to the performance of the LSA summarizer on the negative posts and the second one (Sent+Summ$_{pos}$) presents the performance of the LSA summarizer on the positive posts. The last line contains the results obtained by the best-scoring system in the TAC 2008 summarization track.

System	**R$_1$**	**R$_2$**	**R$_{SU4}$**	**R$_L$**
Sent + Summ$_{neg}$	0.22 (0.18-0.26)	0.09 (0.06-0.11)	0.09 (0.06-0.11)	0.21 (0.17-0.24)
Sent + Summ$_{pos}$	0.21 (0.17-0.26)	0.05 (0.02-0.09)	0.05 (0.02-0.09)	0.19 (0.16-0.23)
Summ$_{TAC08}$	0.348	0.081	0.12	-

Table 5.27: Overall sentiment summarization performance

DISCUSSION OF THE INITIAL RESULTS

The first thing to note from Table 5.27 is that the performance on negative posts is better, though, being within the 95% confidence intervals, the difference cannot be considered statistically significant. One possible reason for the slightly better performance on the negative posts is that the sentiment recognition system is more accurate with negative sentiment than with positive.

The other observation we make is that the results on our corpus are not directly comparable with those of TAC 2008 for two reasons: Firstly, the data sets are different and secondly, we used the F1 score to account better for the variation in size of our model summaries. However, it is worth noting that the LSA summarizer employing the same method as our LSA summarizer ranked in the top 20% summarization systems at the TAC 2008 competition. Additionally, the same LSA method has already been improved upon by incorporating higher level semantic information such as co-reference (Steinberger et al., 2007), and hence, applying the same method in our context would also potentially translate in performance improvement.

In the light of this, we believe the performance results we obtained are promising. The main problem we encountered was that the LSA-based summarization method we adopted was originally designed to work with grammatical sentences from news articles. In our case, however, blog posts are often composed of ungrammatical sentences and, additionally, a high number of unusual combinations of characters such as :-), ;), :-(etc. (corresponding to the so-called \emoticons"), which make the blog data much nosier and harder to process than the standard data sets traditionally used for summarization evaluation. Nevertheless, in our case, the LSA method, being a statistical method, proved to be quite robust to variations in the input data and, most importantly, to the change of domain. We used F1 score instead of recall used at TAC, because the lengths of our model summaries vary from one thread to another.

5.3.3. SUMMARIZATION BASED ON OPINION STRENGTH

INTRODUCTION

Further to our initial experiments (Balahur et al., 2009g), we continued our efforts to develop adequate techniques for opinion summarization (Kabadjov et al., 2009). In the latter analysis by Kabadjov et al. (2009), we explored the impact of sentiment intensity on the summarization performance and, in general, the relationship between these two concepts. In other words, are comments expressing very negative or very positive opinions also salient from the point of view of summarization? Intuitively, sentiment summarization can be different from the summarization of factual data, as sentences regarded as informative from the factual point of view may contain little or no sentiment, so, eventually, they are useless from the sentiment point of view.

The main question we address at this point is: how can one determine, at the same time, both sentiment, as well as information-relevant sentences? In the light of these questions, we proposed adequate methodologies and performed experiments to study the relationship between sentiment intensity and summarization (Kabadjov et al., 2009). This idea was originally discussed by Balahur et al. (2008).

EXPERIMENTAL SETTING

Our approach follows a simple intuition: when people express very negative or very positive sentiment, for example, in blogs, they might be also conveying important and valuable information that is somewhat more salient than other comments. The sub-area of Natural Language Processing concerned with identifying salient information in text documents is Text Summarization, hence, we decided to formalize our intuition in the context of text summarization and make use of standard methodology from that area. In addition, we cast the above intuition as a statistical hypothesis test where the null hypothesis we seek to reject is the opposite of our intuition, that is, the sentiment intensity of salient blog comments is no different from the sentiment intensity of non-salient comments. In order to carry out experiments to study in a quantitative manner whether sentiment intensity is a useful summary indicator, three things are needed: a sentiment analysis system capable of producing a sentiment intensity score for a given blog comment, a summarization algorithm exploiting this sentiment intensity score and a reference corpus annotated for both sentiment and salience (i.e., gold standard data). Next, we describe each of those components and the design of the hypothesis test that we used in our research (Kabadjov et al., 2009).

For the experiments we used the sentiment analysis system described by Balahur et al. (Balahur et al., 2009f). Subsequently, we defined a straightforward summarization algorithm that exploits sentiment intensity in the following manner. The system should:

1. Rank all comments according to their intensity for a given polarity.
2. Select highest-scoring n comments (until the limit in the number of sentences given by the compression rate.

At this stage, it is important to point out that positive and negative polarity comments are treated separately, that is, we produce one summary for all positive comments and one for all negative comments for a given blog thread.

We ran this algorithm at two commonly used compression rates: 15% and 30%. That is, we produce two summaries for each polarity for each thread, one by choosing the top 15% and the other by selecting the top 30% of all comments.

In addition to a standard summarization evaluation, we evaluate the hypothesis that very positive or very negative comments are good choices to be included in a summary, by casting the problem as a statistical hypothesis test. Student's t-test. We define the following setting in order to execute an independent two-sample one-tailed t-test of unequal sample sizes and equal variance:

1. Null hypothesis, $H_0 : \bar{X}_1 - \bar{X}_2 = 0$
   Alternative hypothesis, $H_1 : \bar{X}_1 > \bar{X}_2$
2. Level of significance: $\alpha = 0.05$
3. t statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$S_{X_1 X_2} = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}$$

Criterion: Reject the null hypothesis in favor of the alternative hypothesis if $t > t_{v,\alpha}$ where $v = n_1 + n_2 - 2$ (degrees of freedom) and $t_{\infty,0.05} = 1.645$

In the setting considered, n is the number of sample points, 1 is group one and 2 is group two. More specifically, in our case group one is composed of all the comments annotated as salient in our corpus (i.e., gold summary comments) and group two is composed of all the comments that were not annotated (i.e., gold non-summary comments). Furthermore, we further slice the data upon polarity (as produced by the sentiment analysis tool), so we have two samples (i.e., group one and group two) for the case of positive comments and two samples for the case of negative comments. For example, out of all the comments that were assigned a positive score by the sentiment analysis tool, there are those that were also annotated as positive by the annotators these constitute group one for the positive polarity case and those that were not annotated at all these constitute group two for the positive polarity case. The same thinking applies for the negative polarity case.

EVALUATION OF THE OPINION SUMMARIZATION PROPOSAL BASED ON SENTIMENT INTENSITY

The performance results of the sentiment analysis are shown in Table 5.28.

System	Precision	Recall	F1
Sent_{neg}	0.98	0.54	0.69
Sent_{pos}	0.07	0.69	0.12

Table 5.28: Performance of the sentiment analysis system

The first thing to note in Table 5.28 is that the sentiment analysis tool is doing a much better job at identifying negative comments (F1 = 0.69) than positive ones (F1 = 0.12), the main problem with the latter being a very low precision (P = 0.07). One possible reason for this is an insufficient number of annotated positive examples (there were much more negative examples than positive ones in the corpus). In the next section, we discuss whether this substantial difference in performance between the negative and positive cases has an impact on the subsequent analysis. Performance results of the summarizer are shown in Table 5.29.

System	R_1	R_2	R_{SU4}	R_L
SISumm_{neg} at 15%	0.07	0.03	0.03	0.07
SISumm_{pos} at 15%	0.22	0.03	0.03	0.19
SISumm_{neg} at 30%	0.17	0.06	0.06	0.16
SISumm_{pos} at 30%	0.19	0.03	0.03	0.17
TopSumm_{TAC08}	-	0.111	0.142	-

System	R_1	R_2	R_{SU4}	R_L
$\text{BottomSumm}_{\text{TAC08}}$	-	0.069	0.081	-

Table 5.29: Evaluation of the sentiment summarization system with ROUGE scores

We used the same evaluation metrics as the ones employed in our previous efforts (Balahur et al., 2009g).

There are five rows in Table 5.29: the first ($\text{SISumm}_{\text{neg}}$ at 15%) is the performance of the sentiment-intensity-based summarizer (SISumm) on the negative posts at 15% compression rate; the second ($\text{SISumm}_{\text{pos}}$ at 15%) presents the performance of SISumm on the positive posts at 15% compression rate; the third ($\text{SISumm}_{\text{neg}}$ at 30%) is the performance of the SISumm on the negative posts at 30% compression rate; the fourth ($\text{SISumm}_{\text{pos}}$ at 30%) presents the performance of SISumm on the positive posts at 30% compression rate; and finally, the fifth and the sixth rows correspond to the official scores of the top and bottom performing summarizers at the 2008 Text Analysis Conference Summarization track (TAC08), respectively. The last scores are included to provide some context for the other results. Certainly, in order to use gold polarity alongside the score produced by the sentiment analysis tool as we do, we had to firstly automatically align all the automatically identified sentences with the annotated comments. The criterion for alignment we used was that at least 70% of the words in an automatically identified sentence are contained in an annotated comment for it to inherit the gold polarity of that comment (and by virtue of that to be considered a gold summary sentence).

DISCUSSION

From Table 5.29 it is evident that the ROUGE scores obtained are low (at least in the context of TAC 2008). This suggests that sentiment intensity alone is not a sufficiently representative feature of the importance of comments for summarization purposes. Thus, using it in combination with other features that have proven useful for summarization, such as entities mentioned in a given comment (Balahur et al., 2010d), certain cue phrases and surface features, or features capturing the relevance of blog posts to the main topic, is likely to yield better results. In particular, incorporating topic detection features would be crucial, since at the moment off-topic, but very negative or very positive, comments are clearly bad choices for a summary, and currently we employ no means for filtering these out.

There is also an alternative interpretation of the attained results. These results were obtained by using a methodology used in text summarization research, so it is possible that the method is not particularly well-suited for the task at hand, that of producing sentiment-rich summaries. Hence, the reason for the low results may be that we addressed the problem in the context of a slightly different task, suggesting that the task of producing content-based summaries and that of producing sentiment-based summaries are two distinct tasks which require a different treatment. In addition to the above results, we perform the statistical hypothesis test. The values of the variables and the resulting t-statistic values are shown in Table 5.30.

Polarity	$\bar{X}_1$	$\bar{X}_2$	n_1	n_2	$S^2_{X_1}$	$S^2_{X_2}$	**t statistic**
Negative	-3.95	-4.04	1092	1381	10.13	10.5	**0.021**
Positive	4.37	4.26	48	1268	9.3	28.03	**0.036**

Table 5.30: Values for the variables and resulting t-statistic for the 2-sample t-test, unequal sample sizes, equal variances

In both cases, negative and positive polarity, the t values obtained are not large enough for us to reject the null hypothesis in favor of the alternative hypothesis. That is, we do not have any empirical evidence to reject the null hypothesis that the sentiment intensity of salient blog comments is any different from the sentiment intensity of non-salient comments in favor of our alternative hypothesis that, indeed, sentiment intensity in summary blog comments is different from that of non-summary blog comments.

We conclude that, based on our annotated corpus, the hypothesis that very positive or very negative sentences are also good summary sentences does not hold. But, once again, we point out that these results are meaningful in the context of text summarization, that is, the task of producing content-based summaries. Hence, the observation we made above that producing content based summaries is different from producing sentiment-based summaries and as such these tasks should be treated differently also applies in this case. We note, however, that the results on our corpus are not directly comparable with those of TAC08, since the data sets are different and the tasks involved are significantly distinct. Blog posts in our corpus were annotated as important with respect to the main topic of the respective blog threads.

5.3.4. OPINION SUMMARIZATION USING A TOPIC-SENTIMENT ANALYSIS APPROACH

INTRODUCTION

Subsequently to the initial efforts (Balahur et al., 2009g; Kabadjov et al., 2009), we realized that the sentiment analysis component needs to be enhanced with a topic-detection module, so that the performance of the opinion summarization could increase. These subsequent efforts to include topic detection and determine the sentiment of opinions in a topic-dependent manner were described by Balahur et al. (Balahur et al., 2010e).

As discussed in the preceding sections, the performance of the opinion summarization, as it was tackled so far (without taking into consideration the topic) was rather low.

From the human evaluation of the obtained summaries, we could see that the sentiment analysis system classified the sentences correctly as far as opinion, polarity and intensity are concerned. However, many topic irrelevant sentences were introduced in the summaries, leaving aside the relevant ones. On the other hand, we could notice that in the experiments, taking into consideration the presence of the opinion target and its co-references and computing the opinion polarity around the mentions of the target reaches a higher level of performance.

Therefore, it became clear that first of all, a system performing opinion summarization in blogs must first of all include a topic component. Secondly, the research done so far in this area has not taken into consideration the use of methods to detect sentiment that is directly related to the topic. In the experiments we have performed, we detect sentences where the topic is mentioned, by using Latent Semantic Analysis (LSA). Thirdly, most summarization systems do not take into consideration semantic information or include Named Entity variants and co-references. In our approach, also employed in the TAC 2009 summarization, we employ these methods and show how we can obtain better results through their use.

EXPERIMENTAL SETTING

In the first stage, we employ the same technique as in the preliminary approach, but using only the resources that best scored together (MicroWordNet Opinion, JRC Lists and General Inquirer). We map each of these resources into four classes (positive, negative, high positive and high negative), and assign each of the words in the classes a value, i.e., 1, -1, 4 and -4, respectively. We score each of the blog sentences as sum of the values of the opinion words identified in it. In the second stage, we first filter out the sentences that are associated to the topic discussed, using LSA. Further on, we score the sentences identified as relating to the topic of the blog post, in the same manner as in the previous approach. Subsequently to this, we propose a method to select the topic-related sentences based on LSA.

The aim of this approach is to select for further processing only the sentences which contain opinions on the post topic. In order to filter these sentences in, we first create a small corpus of blog posts on each of the topics included in our collection. These small corpora (30 posts for each of the five topics) are gathered using the search on topic words on http://www.blogniscient.com/ and crawling the resulting pages. For each of these 5 corpora, we apply LSA, using the Infomap NLP Software[52]. Subsequently, we compute the 100 most associated words with two of the terms that are most associated with each of the 5 topics and the 100 most associated words with the topic word. For example, for the term “bank”, which is associated to “economy”, we obtain (the first 20 terms):

```
bank:1.000000;money:0.799950;pump:0.683452;switched:0.682389; interest:0.674177;
easing:0.661366; authorised:0.660222; coaster:0.656544; roller:0.656544;
maintained:0.656216; projected:0.656026; apf:0.655364; requirements:0.650757;
tbills:0.650515; ordering:0.648081; eligible:0.645723;
ferguson's:0.644950;proportionally:0.63358; integrate:0.625096; rates:0.624235
```

The summarization system we employed in these experiments is based on the one described by Steinberger et al. (2009). In this approach, the source representation (that is, the input matrix A to the LSA system) is enriched with semantic information combining several source of knowledge, such as lexical information, information about entities and about hypernymy relationships such as those found in IS-A taxonomies, such as MeSH. For our experiments, we used the NewsExplorer multilingual tools for geo-tagging and entity disambiguation described by Pouliquen et al. (2007) and used them to augment the source entity-by-sentence matrix A used in the LSA-based summarizer proposed by J. Steinberger et al. (2009). In addition, we augmented the matrix with terms grounded to the Medical Subject Headings (MeSH) taxonomy, created by the “Health-On-the-Net” organization. The main idea behind this is to capture more complex semantic relationships such as hypernymy and synonymy. The Medical Subject Headings (MeSH) thesaurus is prepared by the US National Library of Medicine for indexing, cataloguing, and searching for biomedical and health-related information and documents. Although it was initially meant for biomedical and health-related documents, since it represent a large IsA taxonomy, it can be used in more general tasks (http://www.nlm.nih.gov/ mesh/meshhome.html).

EVALUATION RESULTS AND DISCUSSION

For the experimental analysis, we include the usual ROUGE metrics: R1 is the maximum number of co-occurring unigrams, R2 is the maximum number of co-occurring bigrams, RSU4 is the skip bigram measure with the addition of unigrams as counting unit, and finally, RL is the longest common subsequence measure. In the cases of the baseline systems we present the average F1 score for the given metric and within parenthesis the 95% confidence intervals. There are four rows in Table 5.31: the first one, Sent + BLSumm_{neg}, is the performance of the baseline

[52] http://infomap-nlp.sourceforge.net/

LSA summarizer on the negative posts (i.e., using only words), the second one, Sent + Summneg, is the enhanced LSA summarizer exploiting entities and IS-A relationships as given by the MeSH taxonomy, the third one, Sent + BLSumm$_{pos}$, presents the performance of the baseline LSA summarizer on the positive posts and the fourth one, Sent+Summ$_{pos}$, is the enhanced LSA summarizer for the positive posts.

System	R$_1$	R$_2$	R$_{SU4}$	R$_L$
Sent+BLSumm$_{neg}$	0.22 (0.18-0.26)	0.09 (0.06-0.11)	0.09 (0.06-0.11)	0.21 (0.17-0.24)
Sent+Summ$_{neg}$	**0.268**	0.087	0.087	**0.253**
Sent+BLSumm$_{neg}$	0.21 (0.17-0.26)	0.05 (0.02-0.09)	0.05 (0.02-0.09)	0.19 (0.16-0.23)
Sent+Summ$_{neg}$	**0.275**	0.076	0.076	**0.249**

Table 5.31: Results of the opinion summarization process

Based on Table 5.31 we can say that the results obtained with the enhanced LSA summarizer are overall better than the baseline summarizer. The numbers in bold show statistically significant improvement over the baseline system (note they are outside of the confidence intervals of the baseline system). The one exception where there is a slight drop in performance of the enhanced summarizer with respect to the baseline system is in the case of the negative posts for the metrics R2 and RSU4, however, the F1 is still within the confidence intervals of the baseline system, meaning the difference is not statistically significant.

We note that the main improvement in the performance of the enhanced summarizer comes from better precision and either no loss or minimal loss in recall with respect to the baseline system. The improved precision can be attributed, on one hand, to the incorporation of entities and IS-A relationships, but also, on the other hand, to the use of a better sentiment analyzer than the one used to produce the results of the baseline system.

We conclude that by using a combined topic-sentiment approach in opinion mining and exploiting higher-level semantic information, such as entities and IS-A relationships, in the summarization process, we obtain a tangible improvement for the opinion-oriented summarization of blogs.

5.3.5. TOPIC-SENTIMENT ANALYSIS VERSUS SUMMARIZATION AS FIRST STEP IN OPINION SUMMARIZATION

EXPERIMENTAL SETTING

In all our previous approaches, we considered by default that opinion summarization should be done by first employing an opinion mining system and subsequently, a summarizer. In this last experiment, we set out to demonstrate that this order is motivated by the improved quality of the results obtained when performing opinion mining and subsequently summarization, as opposed to firstly applying summarization and secondly opinion mining.

In order to compare the results of these two approaches, we employ the opinion mining system proposed by Balahur et al. (Balahur et al., 2009f) and the summarization system described by Steinberger et al. (2009), on the corpus of bank reviews presented by Saggion and Funk (2010) and also employed by Saggion et al. (2010). Each of the reviews in this set has a number of stars assigned (from 1 to 5, 1 for a very negative opinion and 5 for a very positive one), corresponding to the positive/negative assessment of the bank by the reviewer. There is a total of 89 reviews (17 one-star, 11 two-star, 9 three-star, 28 four-star and 24 five-star). Since our opinion summarization system only uses two classes of sentiment (positive and negative), we exclude from the review set those which are assigned three stars and perform our experiments on the set of 80 reviews where the topic is assessed negatively (1 and 2-star reviews) or positively (4 and 5-star reviews).

Further on, we perform four sets of experiments, designed to evaluate which technique is appropriate for the summarization of opinions, so that in the end, the overall polarity is present in the result.

In the first three experiments, we employ the opinion mining system as the initial step in the process; in the last experiment, we first employ the summarization system and subsequently process the results obtained with the opinion mining system. In these first three experiments, we initially process the reviews with the topic-sentiment opinion mining system described above and compute a sentiment score for each of the sentences in the reviews. Subsequently, in our first experiment, we computed the score of the individual reviews as sum of the scores assigned to the sentences it contains (this is referred to as *Document level* in Table 5.32). In our second approach (referred to as *OM +Top-scoring sentences* in 5.32), we selected the top-scoring 15% of positive and negative sentences in each review and computed the overall score of the review as sum of the normalized positive score (total score of the selected positive sentences divided by the number of selected positive sentences) and the normalized negative score (total score of the selected negative sentences divided by the number of selected sentences). In the third approach (*OM+Summarizer* in Table 5.32), we processed the positive and negative sentences in each review using the summarization system described above and obtained the 15% most salient sentences in each of these two sets. We then computed the overall score of the reviews as sum of the normalized positive score (total score of the selected positive sentences divided by the number of selected positive sentences) and the normalized negative score (total score of the selected negative sentences divided by the number of selected sentences).

In the fourth experiment (*Summarizer+OM* in Table 5.32), we first process each review with the summarization system and obtain the top 15% most important sentences. Subsequently, we compute the sentiment score in each of these sentences, using the opinion mining system.

EVALUATION AND DISCUSSION

The results obtained in these four approaches are presented in Table 5.32. They are also compared against a random baseline, obtained by an average of 10 random baselines computed over a set of 80 examples to be classified into positive or negative (we have 28 negative and 52 positive reviews).

Approach	**Accuracy**
Document level	0.62
OM+Top-scoring sentences	0.76
OM+Summarizer	0.8
Summarizer+OM	0.66
Random baseline	0.47

Table 5.32: Results of opinion summarization for different approaches

As it can be noticed from the results in Table 5.32, performing opinion mining as a prior step to summarization results in a better approximation of the sentiment expressed in the initial text.

The results obtained in the case when summarization is performed prior to sentiment analysis are comparable to the ones obtained when computing the overall sentiment at a document level, showing that the result of summarization does offer a genuine image of what is expressed in the original, but fails to filter in opinion-related information with a higher priority.

Finally, the results show that by employing only the opinion mining system and selecting the top-scoring positive and negative sentences, we outperform the document-level sentiment analysis and summarization as first step approaches, but obtain lower results than in the case of using opinion mining as a first processing step, followed by summarization. We believe that the difference is due to the fact that when scoring the sentiment present in a sentence, its relevance as far as topic is concerned is only taken into account as filtering factor, and has no influence on the sentence score. This shortcoming can be overcome, as seen in the results, by using the summarization system, which adds information on the importance of the sentences as far as the information content is concerned.

5.3.6. CONCLUSIONS ON THE PROPOSED APPROACHES FOR OPINION SUMMARIZATION

In this second part of this chapter, we presented and evaluated different methods for opinion summarization in the context of blogs and blog threads. In these experiments, the aim was to create a robust system that is able to assess mass opinion on different topics and present the main arguments in favor and against them.

Within the given setting, we showed that a mere combination of an opinion mining system with a summarization system is not sufficient to tackle the task. Further on, we showed that opinion summarization is different from content-based summarization. Subsequently, we proposed a method to extend the original approach by integrating topic-opinion analysis and semantic information, achieving significantly better performance. We used an annotated corpus and the standard ROUGE scorer to automatically evaluate the performance of our system. Finally, we assessed the importance of the order in which opinion mining and summarization are applied to texts, so that the final result of the opinion summarization process offers an accurate image of the opinions expressed in the initial document. The different approaches showed that in the case of opinion summarization, performing the summarization step first can lead to the loss of information that is vital from the opinion point of view.

All in all, we have shown that performing traditional tasks in the context of opinionated text has many challenges. In the case of opinion questions, new elements have to be defined (such as Expected Polarity Type, Expected Source, Expected Target), in order for the task to be correctly tackled. In the case of opinion summarization we have shown that the sentiment analysis system must be employed prior to the summarization system and that the sentiment analysis component must be enhanced with topic-detection mechanisms.

Finally, we have shown that in the case of opinionated text, relevance is given not only by the information contained, but also by the polarity of the opinion and its intensity. Although initial results have shown that there is no correlation between the Gold Standard annotations and the intensity level of sentences, as output by the sentiment analysis system, given the fact that using this method, we obtained high results as far as F-measure is concerned in TAC 2008, we believe that more mechanisms for opinion intensity should be studied, so that the clear connection between sentence relevance and the opinion it contains, as well as the intensity it has, can be established.

CHAPTER 6. DISCOVERING IMPLICIT EXPRESSIONS OF SENTIMENT FROM TEXT

Motto: *"There are moments in life, when the heart is so full of emotion/ That if by chance it be shaken, or into its depths like a pebble/Drops some careless word, it overflows, and its secret,/ Spilt on the ground like water, can never be gathered together" (Henry Longfellow)*

6.1. INTRODUCTION

In the previous chapters, we explored the task of sentiment analysis in different text types and languages, proposing a variety of methods that were appropriate for tackling the issues in each particular text type. Most of the times, however, the approaches we took were limited to discovering only the situations where sentiment was expressed explicitly (i.e. where linguistic cues could be found in the text to indicate it contained subjective elements or sentiment).

Nevertheless, in many cases, the emotion underlying the sentiment is not explicitly present in text, but is inferable based on commonsense knowledge (i.e. emotion is not explicitly, but implicitly expressed by the author, by presenting situations which most people, based on commonsense knowledge, associate with an emotion). In this final chapter, we will present our contribution to the issue of automatically detecting emotion expressed in text in an implicit manner.

Firstly, we present our initial approach, which is based on the idea that emotion is triggered by specific concepts, according to their *relevance*, seen in relation to the basic needs and motivations (Maslow, 1943; Max-Neef 1990). This idea is based on the Relevance Theory (Sperber and Wilson, 2000). Subsequently, based on the Appraisal Theory models (De Rivera, 1977; Frijda, 1986; Ortony, Clore and Collins, 1988; Johnson-Laird and Oatley, 1989), we abstract on our initial idea and set up a framework for representing situations described in text as chains of actions and their appraisal values, in the form of a knowledge base. We show the manner in which additional knowledge on the properties of the concepts involved in such situations can be imported from external sources and how such a representation is useful to obtain an accurate label of the emotion expressed in text, without any linguistic clue being present therein.

6.1.1. MOTIVATION

Remembering the definition we provided in Chapter 2, sentiment[53] suggests a settled opinion reflective of one's feelings -"the conscious subjective experience of emotion" (Van den Bos, 2006). Thus, sentiments cannot be present without an emotion being expressed in text, either implicitly, or explicitly. Due to this reason, detecting implicit expressions of emotion can increase the performance of sentiment analysis systems, making them able to spot sentiment even in the cases where it is not directly stated in text, but results as a consequence of the reader's emotion, as a consequence of interpreting what is said.

Detecting emotion is a more difficult task than the mere sentiment analysis from text, as the task includes classification between a larger number of categories (i.e emotion labels), which are not as easily separable or distinguishable as the "positive" and "negative" classes, because of their number (at least 6 basic emotions[54]) and characteristics. Although emotion detection is a related problem to sentiment analysis, we chose to present the work done in this area separately, as we consider that the approaches in the first problem are more difficult and require specific methods and tools to be tackled.

This chapter is structured as follows: we first give a brief introduction on the concepts of emotion and the background of the work presented. Subsequently, we describe and evaluate the "emotion trigger" method, put forward by Balahur and Montoyo (2008), in which the main idea is to create a collection of terms that invoke an emotion based on their relevance to human needs and motivations. Finally, we present EmotiNet – the framework we built for the detection of emotion implicitly expressed in text (Balahur et al, 2011a; Balahur et al., 2011b). The underlying mechanism of this framework is the EmotiNet knowledge base, which was built on the idea of situation appraisal using commonsense knowledge.

6.1.2. THE CONCEPT OF EMOTION

For the reader's convenience, we will first repeat some of the definitions given in Chapter 2 to the term "emotion" and the most related concepts.

Emotion is commonly defined as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems (Information processing, Support, Executive, Action, Monitor) in response to the evaluation of an

[53] http://www.merriam-webster.com/

[54] One of the most widely used classifications of emotions is that of Paul Ekman (1972), which includes 6 basic emotions. Other models, such as the ones proposed by Parrot (2001) or Plutchik (2001) include a higher number of basic emotions, as well as secondary and tertiary emotions.

external or internal stimulus event as relevant to major concerns of the organism" (Scherer, 1987; Scherer, 2001).

The term *feeling* points to a single component denoting the subjective experience process (Scherer, 2005) and is therefore only a small part of an emotion.

Moods are less specific, less intense affective phenomena, product of two dimensions - energy and tension (Thayer, 2001).

Sentiment is "the conscious subjective experience of emotion" (Van den Bos, 2006).

6.1.3. BACKGROUND

Understanding the manner in which humans express, sense and react to emotion has always been a challenge, each period and civilization giving a distinct explanation and interpretation to the diversity of sentiments (Oatley, 2004). Societies used emotions for the definition of social norms, for the detection of anomalies, and even for the explanation of mythical or historical facts (e.g. the anger and wrath of the Greek Gods, the fear of the unknown and the Inquisition in Middle Age, the romantic love in Modern Times) (Ratner, 2000). The cultural representations concerning emotions are generally ambivalent (Goldie, 2000, Evans, 2001, Oatley et al., 2006). Emotions were praised for their persuasive power (Aristotle defines "pathos" – the ability to appeal to the audience's emotions as the second component of the art of rhetorics), but also criticized as a "weakness" of the human being, which should ideally be rational.

Different scientific theories of emotion have been developed along the last century of research in philosophy, psychology, cognitive sciences or neuroscience, each trying to offer an explanation to the diversity of affect phenomena. There were different attempts to build systems that automatically detect emotion from text in the 70s and 80s. However, it was not until 1995, when Rosalind Picard consecrated the term "affect computing" in Artificial Intelligence (Picard, 1995) that the interest computer engineers expressed towards the research in emotion increased significantly. The need to develop systems that are able to detect and respond to affect in an *automatic* manner has become even more obvious in the past two decades, when a multitude of environments of interaction between humans and computers has been built – e.g. e-learning sites, social media applications and intelligent robots. On the one hand, if such environments are able to detect emotion, they can better adapt to the user needs. On the other hand, if they are able to express emotion, they create a more natural type of interaction. Despite the fact that Picard (1995) identified three different types of systems dealing with automatic affect processing (systems detecting emotions, systems expressing what a human would perceive as emotion and systems feeling an emotion), most of the research in affect computing has so far concentrated solely on the first type of systems (Calvo and D'Mello, 2010).

In Natural Language Processing (NLP), the task of detecting emotion expressed in text has grown in importance in the last decade, together with the development of the Web technologies supporting social interaction. Although different approaches to tackle the issue of emotion detection in text have been proposed by NLP researchers, the complexity of the emotional phenomena and the fact that approaches most of the times contemplate only the word level have led to a low performance of the systems implementing this task - e.g. the ones participating in the SemEval 2007 Task No. 14 – (Strapparava and Mihalcea, 2007). The first explanation for these results, supported by linguistic studies and psychological models of emotion, is that expressions of emotion are most of the times not direct, through the use of specific words (e.g. "I am angry."). In fact, according to a linguistic study by Pennebaker et al (2003), only 4% of words carry an affective content. Most of the times, the affect expressed in text results from the interpretation of the situation presented therein (Balahur and Montoyo, 2008; Balahur and Steinberger, 2009), from the properties of the concepts involved and how they are related within the text. In this sense, the first experiments we performed aimed at building a lexicon of terms whose presence in text trigger emotion. Subsequently, we described a framework for detecting and linking concepts (and not just words) that are used to implicitly express emotion in a text.

6.2. EMOTION DETECTION BASED ON "EMOTION TRIGGERS"

6.2.1. DEFINITION OF "EMOTION TRIGGERS"

Most of the systems that we have presented so far base their analysis on the spotting of linguistic cues of sentiment (e.g. finding words such as "good", "bad", "happy", etc.). Nevertheless, there are many examples in which emotions (and hence sentiments) are expressed and elicited in an indirect manner.

In the example "*Government approves new taxes for car imports"*, one reader will remain neutral, another will be infuriated, a car seller might be glad. Or in the case of a news title such as *"The Spanish Civil War now a computer game"*, a Spanish person could feel offended, another outraged, another amused, a person of another nationality might feel neutral and a computer games' addict very happy. On the other hand, a title such as *"Children killed in bomb attack"* will most certainly produce a general feeling of sadness and/or fear to the readers. While in this latter example we can find different words that we can link to sentiment (e.g. "killed", "bomb", "attack"), the first two examples are objective in nature.

In order to determine the emotion elicited by a text that is apparently objective in its statements, in (Balahur and Montoyo, AISB 2008) we introduced the concept of “emotion triggers”. They are defined as a words or concepts expressing an idea, that depending on the reader’s world of interest, cultural, educational and social factors, leads to an emotional interpretation of the text content or not.

Examples of emotion triggers are “freedom”, “salary”, “employment”, “sale”, “pride”, “esteem”, “family” and so on.

In our initial efforts (Balahur and Montoyo, 2008), we described the process of building a lexicon of such emotion triggers, classifying them according to their polarity and integrating them in a system which spots and classifies the polarity of the sentiment and the emotion expressed therein.

6.2.2. THEORIES UNDERLYING THE “EMOTION TRIGGERS” APPROACH

In order to build the lexicon of emotion triggers, we were inspired by three different theories:

1. The Relevance Theory (Sperber and Wilson, 2000) from Pragmatics.
2. Abraham Maslow’s theory of human motivation and its corresponding pyramid of human needs.
3. Max-Neef’s matrix of fundamental human needs.

THE THEORY OF RELEVANCE

The Theory of Relevance, arises from pragmatics, and states in the cognitive principle that “human cognition tends to be geared toward the maximization of relevance”, that is, from the multiple stimuli present in a communication, be it written or spoken, a reader will choose the one with highest significance to their world of interest.

In the case of emotions, motivated by the work done in automatic argumentation analysis (Mazzotta et al., 2008), we considered that the basic stimuli for emotions are related to the basic needs and motivations. Therefore, the core of our lexicon is represented by the terms included in Maslow’s Pyramid of Human Needs and Motivations and Max-Neef’s matrix of fundamental human needs. We explain in detail the components of these two representations.

MASLOW’S PYRAMID OF HUMAN NEEDS AND MOTIVATIONS

Abraham Maslow (1943), classified the human needs and motivational factors into a 5-level pyramid, from the basic, physiological ones, to the more education and personal level of development dependent ones. Needs such as food, shelter, peace are at the bottom of the pyramid, whereas needs for self achievement, fame, glory are at the top. The basic needs are the general human ones; as we move towards the top, we find the more individual dependent ones.

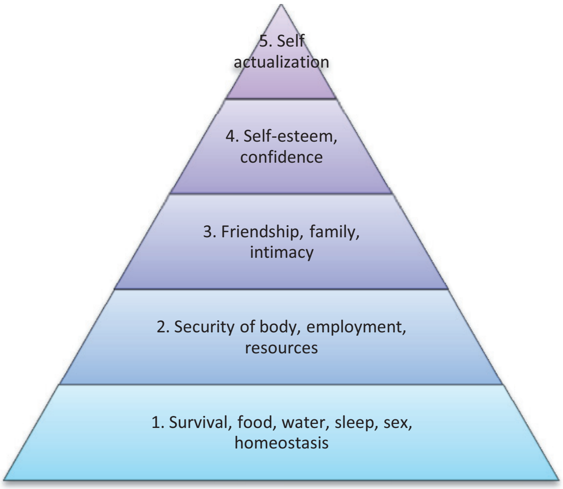

5. Self actualization
4. Self-esteem, confidence
3. Friendship, family, intimacy
2. Security of body, employment, resources
1. Survival, food, water, sleep, sex, homeostasis

Figure 6.1. Maslow's pyramid of human needs and motivations

We consider the terms in Maslow's pyramid levels as primary emotion triggers, very general notions that express ideas that are fundamental to all human beings. In order to exploit this classification, we build a lexical database of emotion triggers at the 5 levels. The words found in the five levels are nouns, verbs, adjectives and adverbs.

MAX-NEEF'S MATRIX OF FUNDAMENTAL HUMAN NEEDS

Among the critics of the Maslow theory of human needs is Manfred Max Neef, in (Max-Neef, 1991). Human needs, according to Neef, are understood as a system - i.e. they are interrelated and interactive. Max-Neef classifies the fundamental human needs as: subsistence, protection, affection, understanding, participation, recreation (in the sense of leisure, time to reflect, or idleness), creation, identity and freedom. Needs are also defined according to the existential categories of being, having, doing and interacting, and from these dimensions, a 36 cell matrix is developed which can be filled with examples of satisfiers for those needs[55].

Therefore, starting in parallel from the matrix of fundamental human needs as primary emotion triggers, we see, on one hand, if a classification of emotion triggers is better than a flat model with rules of inference, and on the other hand, can build a fine-grained taxonomy of terms indicating the precise category in which each type of emotion trigger influences the human affect.

[55] http://www.rainforestinfo.org.au/background/maxneef.htm

BUILDING THE LEXICON OF “EMOTION TRIGGERS”

The core of English emotion triggers is built, at the first stage, of the approximately 37 terms found in Maslow´s pyramid of human needs, structured on 5 levels starting from the terms corresponding to the deficiency needs, found on the four bottom levels and having on top the growth needs terms, of achieving the personal potential, on level 5.

Since most of the words are general notions and their number is relatively small (37), we disambiguate them with the sense numbers they have in WordNet 2.1, in order to ensure that further on, the added words will remain with the intended meaning. For each term, we add all the senses and all grammatical categories that are valid in the context of Maslow´s pyramid levels. We then add to these words the corresponding synonyms and hyponyms from WordNet. For the verbs considered, we also add the entailed actions. We consider as having a negative value the emotion triggers that are antonyms of the nouns found. For each of the nouns and verbs, we further add the corresponding nouns and verbs, respectively, using NomLex (Macleoud, 1998). Since NomLex does not assign sense numbers to distinguish between the possible semantics of the nouns and verbs in the collection, we use the Relevant Domains concept and corresponding repository (Vázquez et al., 2007) to preserve the intended meaning, by taking the top relevant domain of each word sense and assigning the corresponding verb or noun in NomLex the sense number that has the same top relevant domain. If more such senses exist, they are all added.

On the other hand, another core of English words is completed with the terms found in Max Neef´s matrix of fundamental human needs. This matrix is built according to the four main characteristics of the individual: being, having, doing and interacting, for which terms are assigned in order to nine categories of needs: identity, subsistence, affection, creation, protection, freedom, participation, leisure and understanding.

When building the core of words corresponding to the taxonomy proposed by Neef, we start with the terms semantically disambiguated according to the WordNet 2.1 sense numbers. As in the case of the concepts extracted from Maslow’s 5-level pyramid, we then add to these words the corresponding synonyms and hyponyms from WordNet. For the verbs considered, we also add the entailed actions. We consider as having a negative value the emotion triggers that are antonyms of the nouns found. For each of the nouns and verbs, we further add the corresponding nouns and verbs, respectively, using NomLex. Since NomLex does not assign sense numbers to distinguish between the possible semantics of the nouns and verbs in the collection, we use the Relevant domain concept and corresponding repository to preserve the intended meaning, by taking the top relevant domain of each word sense and assigning the corresponding verb or noun in NomLex the sense number that has the same top relevant domain. If more such senses exist, they are all added.

Using EuroWordNet[56], we map the words in the English lexical database of emotion triggers to their Spanish correspondents, preserving the meaning through the WordNet sense numbers.

The final step in building the lexical databases consists of adding real-world situations, cultural-dependent contexts terms to the two lexical databases. For English, we add the concepts in ConceptNet[57] that are linked to the emotion triggers contained so far in the lexicon based on the relations *DefinedAs*, *LocationOf*, *CapableOf*, *PropertyOf* and *UsedFor*. For Spanish, we add the cultural context by using the Larousse Ideologic Dictionary of the Spanish Language.

6.2.4. ASSIGNMENT OF POLARITY AND EMOTION TO THE EMOTION TRIGGERS

The next step consists in assigning polarity and emotion to the terms in the database. This is done with the following rules, both for the terms in Maslow's pyramid as well as for those in Neef's matrix:

1. The primary emotion triggers are assigned a positive value.
2. The terms (also emotion triggers in the final lexical database) synonyms and hyponyms of the primary emotion triggers, as well as the entailed verbs are assigned a positive value.
3. The terms opposed and antonym of those from 1. and 2. are assigned a negative valence.
4. Emotion triggers added further on inherit the valence from the emotion trigger they are related to in case of synonyms, hyponyms and entailment and change their valence from positive to negative or negative to positive in the case of antonyms.
5. Value of all emotion triggers is modified according to the valence shifters they are determined by.

Further on, we assign an emotion triggers a value on each of the 6 categories of emotion proposed for classification in the SemEval Task No. 14 – joy, sadness, anger, fear, disgust and surprise, using the following rules:

1. The emotion triggers found in the levels of Maslow´s pyramid of needs and those found in the components of Neef´s matrix of fundamental human needs are manually annotated with scores for each of the 6 categories.

[56] http://en.wikipedia.org/wiki/EuroWordNet

[57] http://web.media.mit.edu/~hugo/conceptnet/

2. The primary emotion triggers are assigned a positive value.
3. The terms (also emotion triggers in the final lexical database) synonym and hyponym of the primary emotion triggers, as well as the entailed verbs are assigned a positive value.
4. The terms opposed and antonym of those from 1. and 2. are assigned a negative valence.
5. Emotion triggers added further on inherit the valence from the emotion trigger they are related to in case of synonyms, hyponyms and entailment and change their valence from positive to negative or negative to positive in the case of antonyms with opposed values.
6. Value of all emotions of an emotion triggers is modified according to the valence shifters they are determined by.
7. If any of the values calculated in 6 is higher than 100, it is set to 100; if it is lower than -100, it is set to -100.

6.2.5. A METHOD FOR EMOTION DETECTION IN TEXT BASED ON "EMOTION TRIGGERS"

In order to be able to recognize the change in meaning of emotion triggers due to modifiers, we have defined a set of valence shifters – words that negate the emotion triggers, intensify or diminish their sense. The set contains:

- Words that introduce negation (no, never, not, doesn´t, don´t and negated modal verbs)
- A set of adjectives that intensify the meaning of the nouns they modify – big, more, better etc.
- A set of adjectives that diminish the meaning of the nouns they modify – small, less, worse etc.
- The set of modal verbs and conditional of modal verbs that introduce uncertainty to the active verb they determine- can, could, might, should, would
- The set of modal verbs that stress on the meaning of the verb they determine - must
- A set of adverbs that stress the overall valence and intensify emotion of the context – surely, definitely etc
- A set of adverbs that shift the valence and diminish emotion of the context – maybe, possibly etc

For each of the valence shifters, we define a weight of 1.5 for the meaning intensifiers and 0.5 for the meaning diminishers. These are coefficients that will be multiplied with the weight assigned to the emotion trigger level and emotions- level association ratio corresponding to the given emotion trigger in the case of emotion triggers built from Maslow´s pyramid. In the case of emotion triggers stemming from Neef`s matrix of fundamental human needs, the weights of the valence shifters are multiplied with the emotion-category association ratio, computed for each emotion trigger and each of the four existential categories. In order to determine the importance of the concepts to a specific domain, we will employ the association ratio formula.

The association ratio score provides a significance score information of the most relevant and common domain of a word. The formula for calculating it is:

$$AR(w;D) = \Pr(w,D)\log_2 \frac{\Pr(w,D)}{\Pr(w)\Pr(D)}, \text{ where:}$$

- Pr(w,D) is the probability of the word in the given domain
- Pr(w) is the probability of the word
- Pr(D) is the probability of the domain

In our approach, besides quantifying the importance of each emotion trigger in a manner appropriate to the level and emotion it conveys, we propose to use a variant of the association ratio that we call emotion association level. This score will provide the significance information of the most relevant emotion to each level. The corresponding formula is therefore:

$$AR(e;L) = \Pr(e,L)\log_2 \frac{\Pr(e,L)}{\Pr(e)\Pr(L)}, \text{ where:}$$

- Pr(e,L) is the probability of the emotion in the given level
- Pr(e) is the probability of the emotion
- Pr(L) is the probability of the level

The Construction-Integration Model is a psychological model of text comprehension (Kintsch, 1999), based on the idea that while reading a text, a person will activate the features of words that are appropriate to the context and inhibit those that are not.

In this model, the process of text comprehension consists of two phases. The first one – the construction - uses rules in the form a production system to generate, from the linguistic representation of words, a propositional network of related mental elements. Further, adding the knowledge experience of the reader, a more elaborated propositional network is created.

The second phase – integration- takes as input the crude representation of text in the form of the elaborated propositional network, with nodes linked with positive and negative connections meant to represent the relations between them, and tunes it using connectionist relaxation techniques.

6.2.6. IMPLEMENTATION OF THE EMOTION DETECTION METHOD BASED ON “EMOTION TRIGGERS”

The final system built to classify text according to the polarity of the sentiment expressed and the emotion it contained is depicted in Figure 6.2:

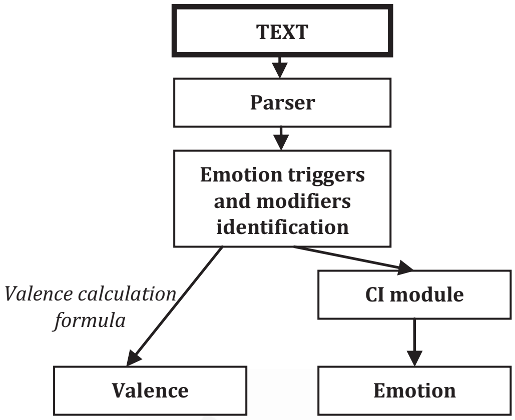

TEXT
Parser
Emotion triggers and modifiers identification
Valence calculation formula
CI module
Valence
Emotion

Figure 6.2. System valence and emotion classification-components

First, the input text is parsed with Minipar to obtain for each word the grammatical category, the lemma and its modifiers. Further on, the emotion triggers in the text are identified, together with their corresponding modifiers.

We calculate the valence of the text on the basis of the identified emotion triggers and their modifiers, using the formulas described in what follows.

In the case of emotion triggers obtained from Maslow´s pyramid, we calculate a score called weighted valence of emotion trigger (wv) using the following formula:

$wv(et_{ij}) = w(m) * w(l_j) * v(et_i)$, where

- w(m) is the weight of modifier
- w(l_j)is the weight of level
- v(et_i)is the emotion trigger valence
- i is the index of the emotion trigger
- j is the number of the level

In the case of emotion triggers obtained from Neef’s matrix, we calculate a score called weighted valence of emotion trigger (wv) using the following formula:

$wv(et_i) = w(m) * v(et_i)$, where

- w(m) is the weight of modifier
- v(et_i)is the emotion trigger valence
- i is the index of the level

The total valence of text equals the sum of all weighted valences of all emotion triggers. The obtained value is rounded to the closest of the two possible values : 0 and 1.

Further on, we calculate the emotions present in the text, by the following method:

- for each emotion trigger stemming from Maslow´s pyramid, we compute the emotion to level association ratio
- for each emotion trigger stemming from Neef´s matrix, we the emotion to category association ratio

We then apply the Construction Integration Model in a manner similar to that described by Lemaire (2005) and construct a spreading activation network. We consider the working memory as being composed of the set of emotion triggers and their emotion association ratio value which is considered as activation value. The semantic memory is set up of the modifiers and the top 5 synonyms and antonyms of emotion triggers with their AR value. We set the value of each emotion trigger to 1. We create a link between all concepts in the semantic memory with all the emotion triggers. We consider the strength of link the higher of the two Emotion trigger Association Ratio scores.

The text is processed in the order in which emotion triggers appear and finally obtain the activation value for each emotion trigger.

The output for the values of the emotions in text is obtained by multiplying the activation values with 100 and adding the scores obtained for the same emotion from different emotion triggers when it is the case.

6.2.7. EVALUATION OF THE EMOTION DETECTION METHOD BASED ON "EMOTION TRIGGERS"

The evaluation of the system presented was done using the test data provided within the SemEval Task No. 14: Affective Text test set (Strapparava and Mihalcea, 2007) and its Spanish translation. In the task proposed in SemEval, the objective was to assign valence – positive or negative - and classify emotion of 1000 news headlines provided as test set according to 6 given emotions: joy, fear, sadness, anger, surprise and disgust. In order to test our emotion trigger method, we employed this test set and its translation to Spanish. The results we obtained are presented in Table 6.1 for valence classification and in Table 6.2 for one of the 6 emotions – fear, which is the predominant emotion in the "emotion triggers" lexicon (for the F measure we considered alpha 0.5):

	Precision	**Recall**	**F-measure**
English	75.23	65.01	69.74
Spanish	71.1	66.13	68.52

Table 6.1. System results for valence annotation

	Precision	**Recall**	**F-measure**
English	47.21	45.37	46.27
Spanish	46.01	43.84	44.89

Table 6.2. System results for emotion annotation for "fear"

Motivated by the low results and the low coverage of the resource, we subsequently proposed a more thorough method to detect emotion from text using commonsense knowledge.

6.3. EMOTION DETECTION FROM TEXT USING APPRAISAL CRITERIA BASED ON COMMONSENSE KNOWLEDGE

6.3.1. INTRODUCTION

The different emotion theories proposed in psychology give various explanations as to why certain episodes lead to a specific affective state: models viewing emotions as expressions believe that there are specific "action tendencies" given by stimuli; models of emotions as embodiments view them as physiological "changes"; models of emotions as social constructs view emotions as built through experience and language; cognitive approaches to emotion – the so-called "appraisal theories" (Scherer, 1999) state that an emotion can only be experienced by a person if it is elicited by an appraisal of an object that directly affects them and that the result is "based on the person's experience, goals and opportunities for action" (Calvo and D'Mello, 2010).

The latter model, besides being the leading one in cognitive psychology, is also the one that can be best applied to detect emotions expressed in text, as most of the factors that explain affect in other theories cannot be extracted from natural language only (the *physiological changes* factor, for example, cannot be detected from text unless they are explicitly stated therein).

The aim of this research is to:

1. Propose a method for modeling affective reaction to real-life situations described in text, based on the psychological model of the appraisal theory. Practically, we propose the modeling of situations presented in text as **action chains** (changes produced by or occurring to an agent related with

the state of a physical or emotional object) and the **context** in which they take place, using ontological representations. In this way, we abstract from the treatment of texts as mere sequences of words to a conceptual representation, able to capture the semantics of the situations described, in the same manner as psychological models claim that humans do.

2. Design and populate a knowledge base of action chains called EmotiNet, based on the proposed model. We will show the manner in which using the EmotiNet ontologies, we can describe the elements of the situation (the actor, the action, the object etc.) and their properties - corresponding to appraisal criteria. Moreover, we demonstrate that the resource can be extended to include all such defined criteria, either by automatic extraction, extension with knowledge from other sources, such as ConceptNet (Liu and Singh, 2004) or VerbOcean (Chklovski and Pantel, 2004), inference or by manual input.

   Motivated by the fact that most of the research in psychology has been made on self-reported affect, the core of the proposed resource is built from a subset of situations and their corresponding emotion labels taken from the International Survey on Emotion Antecedents and Reactions (ISEAR) (Scherer and Walbott, 1997).
3. Propose and validate a method to detect emotion in text based on EmotiNet using new examples from ISEAR. We thus evaluate the usability of the resource and demonstrate the appropriateness of the proposed model.

6.3.2. CONTEXT OF RESEARCH

Affect-related phenomena have traditionally been studied in depth by disciplines such as philosophy or psychology. However, due to the advances in computing and the growing role of technology in everyday life, the past decades have shown an increasing interest in building software systems that automatically process affect. In order for such systems to benefit from the knowledge acquired in social sciences, interdisciplinary methods have been proposed, which use the existing theoretical models as basis for engineering computational ones.

This section explores the state of the art in the three domains our present research is closely related to: approaches to emotion detection in artificial intelligence, appraisal models in psychology and knowledge bases in NLP applications.

EMOTION DETECTION SYSTEMS IN ARTIFICIAL INTELLIGENCE

In Artificial Intelligence (AI), the term *affective computing* was first introduced by Picard (1995). Although there were previous approaches in the 80s and 90s, in the field of NLP, the task of emotion detection has grown in importance together with the exponential increase in the volume of subjective data on the Web in blogs, forums, reviews, etc. Previous approaches to spot affect in text include the use of models simulating human reactions according to their needs and desires (Dyer, 1987), fuzzy logic (Subasic, 2000), lexical affinity based on similarity of contexts – the basis for the construction of WordNet Affect (Strapparava and Valitutti, 2004) or SentiWord-Net (Esuli and Sebastiani, 2005), detection of affective keywords (Riloff et al., 2003) and machine learning using term frequency (Pang et al., 2002; Wiebe and Riloff, 2006). The two latter approaches are the most widely used in emotion detection systems implemented for NLP, because they are easily adaptable across domains and languages. Other proposed methods include the creation of syntactic patterns and rules for cause-effect modelling (Mei Lee et al., 2009). Significantly different proposals for emotion detection in text are given in the work by (Liu et al, 2003) and the recently proposed framework of sentic computing (Cambria et al., 2009), whose scope is to model affective reaction based on commonsense knowledge. Danisman and Alpkocak (2008) proposed an approach based on vectorial representations. The authors compute the set of words that is discriminatory for 5 of the 7 emotions in the ISEAR corpus and represent the examples using measures computed on the basis of these terms.

Finally, an up-to-date survey on the models of affect and their AC applications is presented by Calvo and D’Mello (2010).

APPRAISAL THEORIES

The set of models in psychology known as the appraisal theories claim that emotions are elicited and differentiated on the basis of the subjective evaluation of the personal significance of a situation, object or event (De Rivera, 1977; Frijda, 1986; Ortony, Clore and Collins, 1988; Johnson-Laird and Oatley, 1989). Thus, the nature of the emotional reaction can be best predicted on the basis of the individual’s appraisal of an antecedent situation, object or event. In consequence, there is a need to contextualize emotional response, due to which the same situation can lead to different affective reactions and similar reactions can be obtained through different stimuli.

There are different explanations for the elements considered in the appraisal process (see Scherer, 1999), which are called *appraisal criteria*. Currently, there is no common set of such criteria, as different versions of the appraisal theory have defined their own list of such factors. However, Scherer (1988) shows that the appraisal criteria proposed in the different theories do converge and cover the same type of appraisals.

Examples of such criteria are the ones proposed and empirically evaluated by Lazarus and Smith (1988), organized into a four categories:

i. Intrinsic characteristics of objects and events;

ii. Significance of events to individual needs and goals;

iii. Individual’s ability to cope with the consequences of the event;

iv. Compatibility of event with social or personal standards, norms and values.

Scherer (1988) proposed five different categories of appraisal (novelty, intrinsic pleasantness, goal significance, coping potential, compatibility standard), containing a list of 16 appraisal criteria (suddenness, familiarity, predictability, intrinsic pleasantness, concern relevance, outcome probability, expectation, conduciveness, urgency, cause: agent, cause: motive, control, power, adjustment, external compatibility standards, internal compatibility standards). He later used the values of these criteria in self-reported affect-eliciting situations to construct the vectorial model in the expert system GENESIS (Scherer, 1993). The system maintains a database of 14 emotion vectors (corresponding to 14 emotions), with each vector component representing the quantitative measure associated to the value of an appraisal component. The values for the new situations are obtained by asking the subject a series of 15 questions, from which the values for the appraisal factors considered (components of the vector representing the situation) are extracted. Subsequetnly, the label assigned to the emotional experience is computed by calculating the most similar vector in the database of emotion-eliciting situations.

The appraisal models defined in psychology have also been employed in linguistics. The Appraisal framework (Martin and White, 2005) is a development of work in Systemic Functional Linguistics (Halliday, 1994) and is concerned with interpersonal meaning in text — the negotiation of social relationships by communicating emotion, judgement and appreciation.

KNOWLEDGE BASES FOR NLP APPLICATIONS

As far as knowledge bases are concerned, many NLP applications have been developed using manually created knowledge repositories such as WordNet (Fellbaum, 1998), CYC[58], ConceptNet or SUMO[59]. Some authors tried to learn ontologies and relations automatically, using sources that evolve in time - e.g. Yago

[58] http://cyc.com/cyc/opencyc/overview

[59] http://www.ontologyportal.org/index.html

(Suchanek et al., 2007) which employs Wikipedia to extract concepts, using rules and heuristics based on the Wikipedia categories.

Other approaches to knowledge base population were by (Pantel et al., 2004), and for relation learning (Berland and Charniak, 1999). DIPRE (Brin, 1998) and Snowball (Agichtein and Gravano, 2000) label a small set of instances and create hand-crafted patterns to extract ontology concepts.

6.3.3. ISSUES IN PRESENT APPROACHES

As seen from the previous section, an important body of research already exists in NLP, dealing with emotion detection in text. Thus, it is important to understand why a new approach is needed and in what way it differs and improves the existing ones. We illustrate the need to build a more robust model, starting from a series of examples.

To start from a simple case, a sentence such as (1) “I am happy” should be labeled by an automatic system with “joy”.

Given this sentence, a system working at a *lexical* level would be able to detect the word “happy” (for example using WordNet Affect) and would correctly identify the emotion expressed as “joy”. But already a slightly more complicated example – (2) “I am not happy” – would require the definition of “inverse” emotions and the approach would no longer be straightforward. In the second example, although emotion words are present in the text, additional rules have to be added in order to account for the negation.

Now let us consider the example: (3) “I’m going to a party”, which should be labeled with “joy” as well. A system working at a lexical level would already be overwhelmed, as no word that is directly related to this emotion is present in the text. A method to overcome this issue is proposed in by *sentic computing* (Cambria et al., 2009) and by (Liu et al, 2003), whose main idea is acquiring knowledge on the emotional effect of different concepts. In this manner, the system would know that “going to a party“ is something that produces “joy”. This approach solves the problem of indirectly mentioning an emotion by using the concepts that are relating to it instead. However, it only spots the emotion contained in separated concepts and does not integrate their interaction or context (cognitive or affective) in which they appear. If the example we considered is extended as in (4) “I’m going to a party, although I should study for my exam.”, the emotion expressed is no longer “joy”, but most probably “guilt” (or a mixture of “joy” and “guilt”, but from which guilt prevails). As it can be noticed, even if there are concepts that according to our general knowledge express a certain emotion (e.g. “going to a party”- “joy”), their presence in the text cannot be considered as a mark that the respective sentence directly contains that emotion (e.g. (5) “Going to a party is not always fun.”, which can be a view expressed in a text). Also, their meaning can be completely changed depending on the context in which they appear (e.g. (6) “I must go to this party”, in which the *obligation* aspect completely changes the emotion label of the situation). As we can see from examples (3) to (6), even systems employing world knowledge (concepts instead of words) would fail in most of these cases.

Similarly, we can show that while the fuzzy models of emotion perform well for a series of cases that fit the described patterns, they remain weak at the time of acquiring, combining and using new information.

The most widely used methods of affect detection in NLP are the based on machine learning models built from corpora. Even such models, while possibly strong in determining lexical or syntactic patterns of emotion, are limited as far as the extraction of the text meaning is concerned, even when deep text understanding features are used. They are also limited at the time of, for example, semantically combining the meaning of different situations that taken separately lead to no affective reaction, but their interaction does (i.e. when world knowledge is required to infer the overall emotional meaning of the situation).

Besides the identified shortcomings, which can be overcome by using existing methods, there are also other issues, which none of the present approaches consider. Even if we follow only our own intuition, without regarding any scientific model of emotion, we can say that, for example, the fact that the development of emotional states is also highly dependent on the affect at the current moment; also the context in which the action takes place, the characteristics of the agent performing it, or of the object of the action, and all the peculiarities concerning these elements can influence the emotion felt in a specific situation.

Given the identified pitfalls of the current systems and their impossibility to take into account such factors as context and characteristics of the elements in it, we propose a new framework for modeling affect, that is robust and flexible and that is based on the most widely used model of emotion – that of the appraisal theories.

The implementation of such models showed promising results (Scherer, 1993). However, they simply represented in a quantitative manner the appraisal criteria in a self-reported affective situation, using multiple choice questionnaires. The problem becomes much more complex, if not impossible, when such factors have to be automatically extracted from text. If the appraisal criteria for the actor/action/object of the situation are not presented in the text, they cannot be extracted from it.

Given all these considerations, our contribution relies in proposing and implementing a resource for modeling affect based on the appraisal theory, that can support:

a) The automatic processing of texts to extract:

- The components of the situation presented (which we denote by “action chains”) and their relation (temporal, causal etc.)

- The elements on which the appraisal is done in each action of the chain (agent, action, object);
- The appraisal criteria that can automatically be determined from the text (modifiers of the action, actor, object in each action chain);

b) The inference on the value of the appraisal criteria, extracted from external knowledge sources (characteristics of the actor, action, object or their modifiers that are inferable from text based on common-sense knowledge);

c) The manual input of appraisal criteria of a specific situation.

6.3.4 A METHOD FOR MODELING AFFECTIVE REACTION USING APPRAISAL MODELS

As we have seen, different criteria have been defined for the appraisal process. They can easily be extracted, as in the case of the GENESIS system, when specific questions are asked about them. However, automatically determining all the appraisal criteria from a text is not a trivial issue. Sometimes, this is impossible when these factors are not present in the text, either direcly or inferable from common-sense knowledge - e.g. *familiarity, concern relevance*, *outcome probability, predictability, expectation*. To illustrate this case, we will employ an example of self-reported affective situation from the ISEAR databank and try to answer the questions asked by the GENESIS system based on the information present in the text. Should we be able to extract the answers from the text manually, it would also be possible to perform this extraction process automatically. Table 6.3 presents a self-reported situation when the emotion "joy" was felt (more information on the person recalling this situation can be obtained from the ISEAR database – e.g. age, sex, religion, etc., but we will only try to answer the questions using the information present in the text). The questions are the ones asked in the GENESIS system, and are taken from (Scherer, 1993). In the third column of the table, we annotate whether or not we can answer the question, using YES (we can answer it based on the information in the text), YES/I (we can answer the question based on inference on the information presented in the text), WK (we can answer it based on our world knowledge), FD (we can answer it based on factual data in the ISEAR database, on the subject describing the experience) and NO (we cannot answer this question).

	Situation	
	I went to buy a bicycle with my father. When I wanted to pay, my father took his purse and payed.	
Q. No.	**Question**	**Can we answer?**
1.	Did the situation that elicited your emotion happen very suddenly or abruptly?	YES
2.	Did the situation concern an event or action that had happened in the past, that had just happened or that was expected in the future?	YES
3.	This type of event, independent of your personal evaluation, would it be generally considered as pleasant or unpleasant?	WK
4.	Was the event relevant for your general well-being, for urgent needs you felt, or for specific goals you were pursuing at the time?	WK
5.	Did you expect the event and its consequences before the situation actually happened?	YES
6.	Did the event help you or hinder you in satisfying your needs, in pursuing your plans or in attaining your goals?	YES
7.	Did you feel that action on your part was urgently required to cope with the event and its consequences?	WK
8.	Was the event caused by your own actions – in other words, were you partially or fully responsible for what happened?	YES/I
9.	Was the event caused by one or several other persons – in other words, were other people fully or partially responsible for what happened?	YES
10.	Was the event mainly due to chance?	NO
11.	Can the occurrence and the consequences of this type of event generally be controlled or modified by human action?	NO
12.	Did you feel that you had enough power to cope with the event – i.e. being able to influence what was happening or to modify the consequences?	NO
13.	Did you feel that, after you used all your means of intervention, you could live with the situation and adapt to the consequences?	WK
14.	Would the large majority of people consider what happened to be quite in accordance with social norms and morally acceptable?	WK/FD
15.	If you were personally responsible for what happened, did your action correspond to your self image?	NO

Table 6.3. Analysis of the possibility to extract answers concerning appraisal criteria from a self-reported affective situation (the questions are reproduced from Scherer, 1993)

As we can see from Table 6.3, the majority of appraisal criteria cannot be extracted automatically from the text, as there is no information on them directly mentioned therein. Some criteria can only be inferred from what is said in the text, others depend on our use of the world knowledge and there are even questions to which we cannot answer, since those details are specific to the person living the reported situation.

Nonetheless, this analysis can offer us a very good insight on the phenomena that is involved in the appraisal process, from which we can extract a simpler representation. Viewed in a simpler manner, a situation is presented as a chain of actions, each with an author and an object; the appraisal depends on the temporal and causal relationship between them, on the characteristics of the actors involved in the action and on the object of the action.

Given this insight, the general idea behind our approach is to model situations as chains of actions and their corresponding emotional effect using an ontological representation. According to the definition provided by Studer et al. (1998), an ontology captures knowledge shared by a community that can be easily sharable with other communities. These two characteristics are especially relevant if we want the recall of our approach to be increased. Knowledge managed in our approach has to be shared by a large community and it also needs to be fed by heterogeneous sources of common knowledge to avoid uncertainties. However, specific assertions can be introduced to account for the specificities of individuals or contexts.

In this manner, we can model the interaction of different events in the context in which they take place and add inference mechanisms to extract knowledge that is not explicitly present in the text. We can also include knowledge on the appraisal criteria relating to different concepts found in other ontologies and knowledge bases (e.g. “The man killed the mosquito.” does not produce the same emotional effect as “The man killed his wife.” or “The man killed the burglar in self-defence.”, because the criteria used to describe them are very different).

At the same time, we can define the properties of emotions and how they combine. Such an approach can account for the differences in interpretation, as the specific knowledge on the individual beliefs or preferences can be easily added as action chains or affective appraisals (properties) of concepts.

6.3.5 BUILDING THE EMOTINET KNOWLEDGE BASE

Based on the model we proposed, we aim at representing chains of actions and their corresponding emotional labels from several situations in such a way that we will be able to extract general patterns of appraisal.

The approach we propose defines a new knowledge base to store action chains, called EmotiNet, which aims to be a resource for detecting emotions in text, and a (semi)automatic, iterative process to build it, which is based on existing knowledge from different sources, mainly text corpora. This process principally aims at extracting the action chains from a document and adding them to the knowledge base.

From a more practical viewpoint, our approach defines an action chain as a sequence of action links, or simply actions that trigger an emotion on an actor. Each specific action link can be described with a tuple (actor, action type, patient, emotional reaction). Specifically, the process proposed was divided into nine steps (between brackets, we specify the manner in which the step was performed and the tools used):

1. Selection of initial set of examples of situations corresponding to the seven emotions from the ISEAR databank –corresponding to the core of the knowledge base (automatic selection from the database, using as filtering condition the mention of a family member);
2. Clustering of examples based on their similarity – used to represent classes of examples that are similar (automatically, using the Lesk similarity implemented in Pedersen's Statistics package and performing clustering with K-Means using Weka[60]);
3. Selection of cluster representatives to be modelled (randomly);
4. Semantic role identification in the examples – to extract triples "subject-action-object" to be added to the KB (using the SRL system in (Moreda et al., 2007));
5. Modeling of the situations based on the "subject – action – object - emotional reaction" model (core examples are manually represented under this tuple form);
6. Modeling of emotions and their interaction based on psychological theories (manually);
7. Evaluation of the obtained ontology (automatically, through consistence verification);
8. Extension of the ontology using the VerbOcean resource (automatically).

ISEAR – SELF-REPORTED AFFECT

Self-reported affect is the most commonly used paradigm in psychology to study the relationship between the emotional reaction and the appraisal preceding it (Scherer, 1999). The affective response to events depends on the context in which they take place. Clues on the affective state of a person, as well as their personal traits influence the emotional reaction to a situation.

[60] http://www.cs.waikato.ac.nz/ml/weka/

In the International Survey of Emotional Antecedents and Reactions (ISEAR)[61] – (Scherer and Wallbott, 1997), the student respondents, both psychologists and non-psychologists, were asked to report situations in which they had experienced all of 7 major emotions (joy, fear, anger, sadness, disgust, shame, and guilt). In each case, the questions covered the way they had appraised the situation and how they reacted. Some examples of entries in the ISEAR databank are: "I felt anger when I had been obviously unjustly treated and had no possibility to prove they were wrong." "A bus drove over my right leg. The event itself was not very frightening but when I had to wait in the emergency ward for three hours and then my leg began to swell, I was frightened." Each example is attached to one single emotion.

In order to have a homogenous starting base, we selected from the 7667 examples in the ISEAR database only the ones that contained descriptions of situations between family members. This resulted in a total of 174 examples of situations where anger was the emotion felt, 87 examples for disgust, 110 examples for fear, 223 for guilt, 76 for joy, 292 for sadness and 119 for shame.

Subsequently, the examples were POS-Tagged using TreeTagger. Within each category, we then computed the similarity of the examples with one another, using the implementation of the Lesk distance in Ted Pedersen's Similarity Package. This score is used to split the examples in each emotion class into six clusters using the Simple K-Means implementation in Weka. The idea behind this approach is to group examples that are similar, in vocabulary and structure. This fact was confirmed by the output of the clusters.

SEMANTIC ROLE LABELING OF SITUATIONS

The next step performed was extracting, from each of the examples, the actions that are described. In order to do this, we employ the semantic role labeling (SRL) system introduced by Moreda et al. (2007). From the output of this system, we manually extract the agent, the verb and the patient (the surface object of the verb). For example, if we use the input sentence "I'm going to a family party because my mother obliges me to", the system extracts two triples with the main actors of the sentences: (I, go, family party) and (mother, oblige, me), related by the causal adverb "because".

Moreover, the subjects of these sentences might include anaphoric expressions or, even, self-references (references to the speaker, e.g. 'I', 'me', 'myself') that have to be solved before creating the instances. The resolution of anaphoric

[61] http://www.unige.ch/fapse/emotion/databanks/isear.html

expressions (not self-references) was accomplished automatically, using a heuristic selection of the family member mentioned in the text that is closest to the anaphoric reference and whose properties (gender, number) are compatible with the ones of the reference. Self-references were also solved (replacement of the mentions of "I" with the speaker), by taking into consideration the entities mentioned in the sentence and deducing the possible relations. In case of ambiguity, we chose the youngest, female (if any) member. Following the last example, the subject of the action would be assigned to the daughter of the family because the sentence is talking about her mother and these triples would be updated: (daughter, go, family_party), (mother, oblige, daughter) and (daughter, feel, angry).

Finally, the action links (triples) were grouped and sorted in action chains. This process of sorting was determined by the adverbial expressions that appear within the sentence, which actually specify the position of each action on a temporal line (e.g. "although" "because", "when"). We defined rules according to which the actions introduced by these modifiers happen prior to or after the current context.

MODELS OF EMOTION

Representing emotions is a challenging task. It has been argued that emotional representations cannot be separated from the experiences they correspond to, as there are no real "labels" that can be assigned to emotions. Other issues in emotion representation are that they cannot be studied individually, they do not happen "instantaneously" (but there is a continuum of emotions). In order to describe the emotions and the way they relate and compose, we employ Robert Plutchik's wheel of emotion (Plutchik, 2001) and Parrot's tree-structured list of emotions (Parrot, 2001). These models are the ones that best overlap with the emotions comprised in the ISEAR databank. Moreover, they contain an explicit modeling of the relations between the different emotions. Plutchik's wheel of emotions contains 8 basic emotions and a set of advanced, composed emotions. The model described by Parrot comprises primary, secondary and tertiary emotions. The primary ones are love, joy, surprise, anger and fear.

Our approach combines both models by adding the primary emotions missing in the first model and adding the secondary and tertiary emotions as combinations of the basic ones. Using this combined model as a reference, we manually assigned one of the seven most basic emotions (*anger, fear, disgust, shame, sadness, joy or guilt*) or the *neutral* value to all the action links obtained after the SRL of examples, thus generating 4-tuples *(subject, action, object, emotion)*, e.g. (daughter, go, family party, neutral) or (mother, oblige, daughter, disgust), that have the appropriate structure to be integrated in the core of EmotiNet.

ONTOLOGY DESIGN

The process of building the core of the EmotiNet knowledge base (KB) of action chains started with the design of the core of knowledge, in our case ontology. Specifically, the design process was divided in three stages:

1. **Establishing the scope and purpose of the ontology**. The ontology we propose has to be capable of defining the concepts required in a general manner, which will allow it to be expanded and specialized by external knowledge sources. Specifically, the EmotiNet ontology needs to capture and manage knowledge from three domains: kinship membership, emotions (and their relations) and actions (characteristics and relations between them).
2. **Reusing knowledge from existing ontologies**. In a second stage, we searched for other ontologies on the Web that contained concepts related to the knowledge cores we needed. At the end of the process, we located two ontologies that would be the basis of our ontological representation: the ReiAction ontology[62], which represents actions between entities in a general manner and whose RDF (Resource Description Framework) graph is depicted in Figure 6.3, and the family relations ontology[63], which contained knowledge about family members and the relations between them.
3. **Building our own knowledge core from the ontologies imported.** This third stage involved the design of the last remaining core, i.e. emotion, and the combination of the different knowledge sources into a single ontology: EmotiNet. In this case, we designed a new knowledge core from scratch based on a combination of the models of emotion presented (see Figure 6.4). This knowledge core includes different types of relations between emotions and a collection of specific instances of emotion (e.g. anger, fear, joy). In the last step, these three cores were combined using new classes and relations between the existing members of these ontologies.

ONTOLOGY EXTENSION AND POPULATION

The next process extended EmotiNet with new types of action and instances of action chains using real examples from the ISEAR corpus. We began the process by manually sellecting a subset of 175 documents from the collection after applying the SRL system proposed by Moreda et al. (2007), with expressions related to all the emotions: anger (25), disgust (25), guilt (25), fear (25), sadness (25), joy (25) and shame (25). The criteria for selecting this subset were the simplicity of the sentences and the variety of actions described.

[62] www.cs.umbc.edu/~lkagal1/rei/ontologies/ReiAction.owl

[63] www.dlsi.ua.es/~jesusmhc/EmotiNet/family.owl

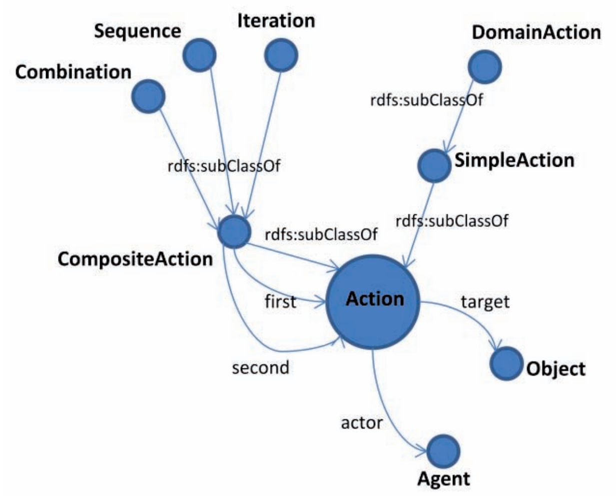


Figure 6.3. Partial RDF graph of ReiAction ontology

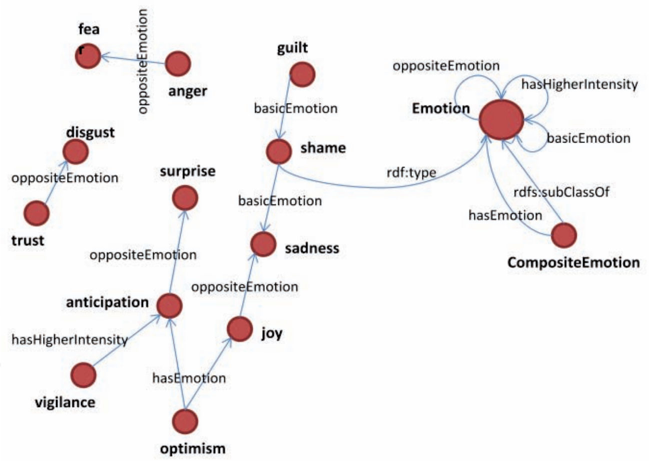


Figure 6.4. Partial RDF graph of the Emotion Ontology.

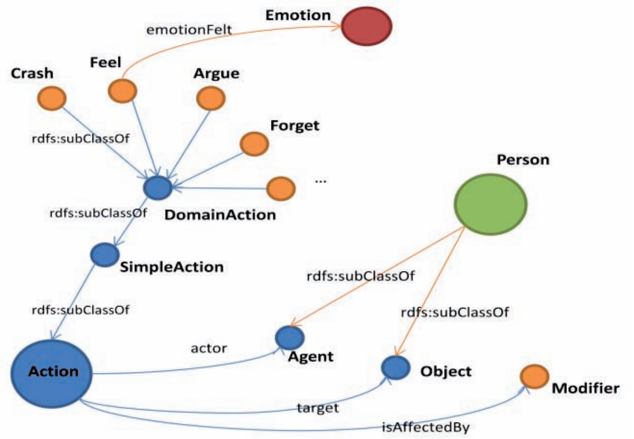


Figure 6.5. Main concepts of EmotiNet.

Once we extracted the actors, actions and objects from the sentences in the 175 situations chosen to be modeled, we ordered them and assigned each of the actions an emotion. Thus, we obtained 175 action chains (ordered lists of tuples). In order to be included in the EmotiNet knowledge base, all their elements needed to be mapped to existing concepts or instances within the KB. When these did not exist, they were added to it.

In EmotiNet, each action link was modeled as a “4-tuple” of ontology instances (actor, action, patient, emotion) that could be affected by a set of modifiers (e.g. opposite, degree of intensity) that directly affects each action. We would like to highlight that each tuple extracted from the process of semantic role labelling and emotion assignation has its own representation in EmotiNet as an instance of the subclasses of Action. Each instance of Action is related to an instance of the class Feel, which represent the emotion felt in this action. Subsequently, these instances (action links) were grouped in sequences of actions (class Sequence) ended by an instance of the class Feel, which, as just mentioned, would determine the final emotion felt by the main actor(s) of the chain.

In our example, we created two new classes Go and Oblige (subclasses of DomainAction) and two new instances from them: instance ”act1”: (class “Go”, actor “daughter”, target “family_party”); and instance “act2”: (class “Oblige”, actor “mother”, target “daughter”). The last action link already existed within EmotiNet from another chain so we reused it: instance “act3”: (class Feel, actor “daughter” emotionFelt “anger”).

The next step consisted in sorting and grouping these instances into sequences by means of instances of the class Sequence. This is a subclass of Action that can establish the temporal order between two actions (which one occurred first). Considering a set of default rules that modelled the semantics of the temporal adverbial expressions, we group and order the instances of action in sequences, which must end with an instance of Feel. Figure 6.6 shows an example of a RDF graph, previously simplified, with the action chain of our example.

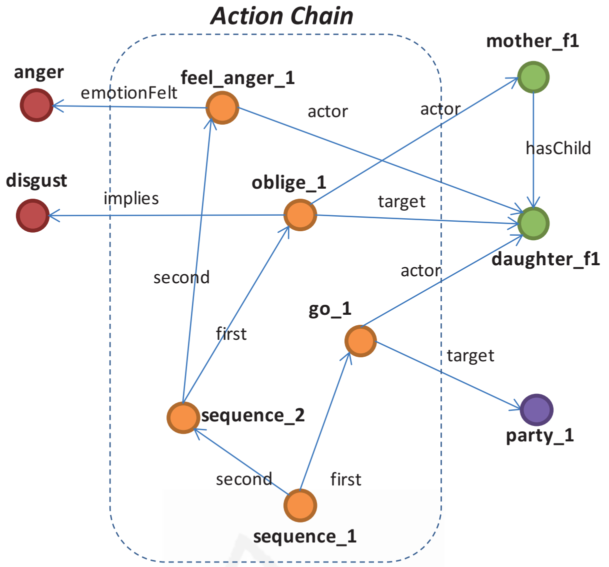


Figure 6.6. RDF graph of an action chain.

Following this strategy, we finally obtained a tight net of ontology instances that express different emotions and how actions triggered them. We used Jena[64] and MySQL for managing and storing the knowledge of EmotiNet on a database in order to carry out our experiments.

ONTOLOGY EXPANSION

Although our knowledge core was favorably assessed, the number of action types was small taking into consideration the number of verbs that appear, for instance, in an English dictionary. In order to extend the coverage of the resource and include certain types of interactions between actions, we expanded the ontology with the actions and relations from VerbOcean (Chklovski and Pantel, 2004). In particular, 299 new actions were automatically included as subclasses of DomainAction, which were directly related to any of the actions of our ontology through three new relations: can-result-in, happens-before and similar. This knowledge extracted from VerbOcean is the basis of inferences when the information extracted from new texts does not appear in our initial set of instances. Thus, this process of expansion is essential for EmotiNet, since it adds new types of action and relations between actions, which might not have analyzed before. This reduced the degree of dependency between the resource and the initial set of examples. The more external

[64] http://jena.sourceforge.net/

sources of general knowledge we add the more flexible will be EmotiNet, thus increasing the possibilities of processing unseen action chains.

6.3.6 EXPERIMENTS AND EVALUATION

The evaluation of our approach consists in testing if by employing the model we built, we are able to detect the emotion expressed in other examples pertaining to the categories in ISEAR.

EXPERIMENTAL SETTINGS

In order to assess the new examples from ISEAR, we followed the same process we used for building the core of EmotiNet, with the exception that the manual modeling of examples into tuples was replaced with the automatic extraction of (actor, verb, patient) triples from the output given by the (Moreda et al., 2007) SRL system. Subsequently, we eliminated the stopwords in the phrases contained in these three roles and performed a simple coreference resolution. Next, we order the actions presented in the phrase, using the adverbs that connect the sentences, through the use of patters (temporal, causal etc.). The resulted action chains represent the test set which will be used in carrying different experiments.

During the development of EmotiNet, we stored two partial versions in order to evaluate the manner in which the number of examples manually included in the EmotiNet core and the number of emotions considered influence the results of the emotion detection task.

The first version of the test set (marked with A) contains as core knowledge on 4 emotions (the number of examples included is specified in brackets): anger (26), disgust (10), guilt (24) and fear (16). The second test set (marked with B) contains knowledge on the same 4 emotions as in A, but with an equal number of examples (25). The third test set (marked with C) comprises the final version of the EmotiNet core knowledge base – containing all 7 emotions in ISEAR, and 25 examples for each.

On all these test sets, we perform the following series of experiments:

(1). In the first experiment, for each of the situations in the test sets (represented as action chains), we search the EmotiNet KB to encounter the sequences in which these actions in the chains are involved and their corresponding subjects. As a result of the search process, we obtain the emotion label corresponding to the new situation and the subject of the emotion based on a weighting function. This function takes into consideration the number of actions and the position in which they appear in the sequence contained in EmotiNet. The issue in this first approach is that many of the examples cannot be classified, as the knowledge they contain is not present in the ontology. The experiment was applied to each of the three test sets (A, B, C), and the corresponding results are marked with A1, B1 and C1, respectively.

(2). A subsequent approach aimed at surpassing the issues raised by the missing knowledge in EmotiNet. In a first approximation, we aimed at introducing extra knowledge from VerbOcean, by adding the verbs that were similar to the ones in the core examples (represented in VerbOcean through the “similar” relation). Subsequently, each of the actions in the examples to be classified that was not already contained in EmotiNet, was sought in VerbOcean. In case one of the similar actions was already contained in the KB, the actions were considered equivalent. Further on, each action was associated with an emotion, using ConceptNet relations and concepts. Action chains were represented as chains of actions with their associated emotion. Finally, new examples were matched against chains of actions containing the same emotions, in the same order. While more complete than the first approximation, this approach was also affected by lack of knowledge about the emotional content of actions. To overcome this issue, we proposed two heuristics:

(2a) In the first one, actions on which no affect information was available, were sought in within the examples already introduced in the EmotiNet and were assigned the most frequent class of emotion labeling them. The experiment was applied to each of the three test sets (A, B, C), and the corresponding results are marked with A2a, B2a and C2a, respectively.

(2b) In the second approximation, we used the most frequent emotion associated to the known links of a chain, whose individual emotions were obtained from SentiWordNet. In this case, the core of action chains is not involved in the process. The experiment was applied to each of the three test sets (A, B, C), and the corresponding results are marked with A2b, B2b and C2b, respectively.

EVALUATION RESULTS

a) Test set A:

We performed the steps described on the 516 examples (ISEAR phrases corresponding to the four emotions modelled, from which the examples used as core of EmotiNet were removed). For the first approach, the queries led to a result only in the case of 199, for the second approach, approximations (A2a) and (A2b) - 409. For the remaining ones, the knowledge stored in the KB is not sufficient, so that the appropriate action chain can be extracted. Table 6.4 presents the results of the evaluations on the subset of examples, whose corresponding query returned a result.

Table 6.5 reports on the recall obtained when testing on all examples. The baseline is random, computed as average of 10 random generations of classes for all classified examples.

b) Test set B:

We performed the steps described on the 487 examples (ISEAR phrases corresponding to the four emotions modelled, from which the examples used as core of EmotiNet were removed). For the first approach, the queries led to a result only in the case of 90, for the second approach, approximations (B2a)- 165 and (B2b) - 171. For the remaining ones, the knowledge stored in the KB is not sufficient, so that the appropriate action chain can be extracted. Table 6.6 presents the results of the evaluations on the subset of examples, whose corresponding query returned a result. Table 6.7 reports on the recall obtained when testing on all examples.

The baseline is random, computed as average of 10 random generations of classes for all classified examples.

c) Test set C:

We performed the steps described on the 895 examples (ISEAR phrases corresponding to the seven emotions modelled, from which the examples used as core of EmotiNet were removed). For the first approach, the queries led to a result only in the case of 571, for the second approach, approximations (C2a) – 617 and (C2b) - 625. For the remaining ones, the knowledge stored in the KB is not sufficient, so that the appropriate action chain can be extracted. Table 6.8 presents the results of the evaluations on the subset of examples, whose corresponding query returned a result. Table 6.9 reports on the recall obtained when testing on all examples.

The baseline is random, computed as average of 10 random generations of classes for all classified examples.

Table 6.10 reproduces the results reported in Danisman and Alpkocak (2008) – which we will mark as DA-, on the ISEAR corpus, using ten-fold cross-validation. We compare them to the results we obtained in the presented experiments. As the authors only present the mean accuracy obtained for 5 of the emotions in ISEAR (anger, disgust, fear, joy, sadness) and they perform ten-fold cross-validation on all examples in the abovementioned corpus, the results are not directly comparable. In fact, a ten-fold cross validation means that they have used 90% of the cases to train the classifier and only tested on the rest of 10% of the cases. As a proof of the significance of the results obtained, we also include the evaluation outcome of the same system on another corpus (marked as DA- SemEval). Again, we cannot directly compare the results, but we can notice that our system performs much better in terms of accuracy and recall when tested on new data.

In any case, we believe that such comparisons can give a clearer idea of the task difficulty and the measure of the success of our approach. On the last line, we include the results reported for the GENESIS system (Scherer, 1993). However, it should be noted that this expert system does not directly detect and classify emotion from text; it only represents answers to a set of questions aimed at determining the values of the appraisal factors included in the emotional episode, after which it computes the similarity to previously computed vectors of situations.

Emotion	**Correct**			**Total**			**Accuracy**		
	A1	**A2a**	**A2b**	**A1**	**A2a**	**A2b**	**A1**	**A2a**	**A2b**
disgust	11	25	25	26	59	63	42.3	42.4	39.7
anger	38	27	26	62	113	113	61.3	23.9	23
fear	4	5	7	29	71	73	16	7.1	9.6
guilt	13	30	26	86	166	160	15.1	18.1	16.3
Total	66	87	84	199	409	409	31.2	21.3	20.5
Baseline	61	84	84	229	409	409	21.9	20.5	20.5

Table 6.4. Results of the emotion detection using EmotiNet on classified examples in test set A

Emotion	**Correct**			**Total**	**Recall**		
	A1	**A2a**	**A2b**	**A1**	**A1**	**A2a**	**A2b**
disgust	11	25	25	76	16.3	32.9	32.9
anger	38	27	26	148	27	18.3	17.6
fear	4	5	7	94	4.5	5.3	7.5
guilt	13	30	26	198	7.7	15.2	13.1
Total	66	87	84	516	14	16.9	16.2
Baseline	112	112	112	516	21.7	21.7	21.7

Table 6.5. Results of the emotion detection using EmotiNet on all test examples in test set A

Emotion	**Correct**			**Total**			**Accuracy**		
	B1	**B2a**	**B2b**	**B1**	**B2a**	**B2b**	**B1**	**B2a**	**B2b**
disgust	10	28	29	41	52	67	24.39	53.85	43.28
anger	16	39	39	102	114	119	15.69	34.21	32.77
fear	37	43	44	55	74	76	67.27	58.11	57.89

Emotion	Correct			Total			Accuracy		
	B1	B2a	B2b	B1	B 2a	B2b	B1	B2a	B2b
guilt	27	55	59	146	157	165	18.49	35.03	35.76
Total	90	165	171	344	397	427	26.16	41.56	40.05

Table 6.6. Results of the emotion detection using EmotiNet on classified examples in test set B

Emotion	Correct			Total	Recall		
	B1	B2a	B2b	B1	B1	B2a	B2b
disgust	10	28	29	59	16.95	47.46	49.15
anger	16	39	39	145	11.03	26.90	26.90
fear	37	43	44	85	43.53	50.59	51.76
guilt	27	55	59	198	13.64	27.78	29.80
Total	90	165	171	487	18.48	33.88	35.11
Baseline	124	124	124	487	0.25	0.25	0.25

Table 6.7. Results of the emotion detection using EmotiNet on all test examples in test set B

Emotion	Correct			Total			Accuracy		
	C1	C2a	C2b	C1	C 2a	C2b	C1	C2a	C2b
disgust	16	16	21	44	42	40	36.36	38.09	52.50
shame	25	25	26	70	78	73	35.71	32.05	35.62
anger	31	47	57	105	115	121	29.52	40.86	47.11
fear	35	34	37	58	65	60	60.34	52.30	61.67
sadness	46	45	41	111	123	125	41.44	36.58	32.80
joy	13	16	18	25	29	35	52	55.17	51.43
guilt	59	68	64	158	165	171	37.34	41.21	37.43
Total	225	251	264	571	617	625	39.40	40.68	42.24

Table 6.8. Results of the emotion detection using EmotiNet on classified examples in test set C

Emotion	Correct			Total	Recall		
	C1	C2a	C2b	C1	C1	C2a	C2b
disgust	16	16	21	59	27.11	27.11	35.59
shame	25	25	26	91	27.47	27.47	28.57
anger	31	47	57	145	21.37	32.41	39.31
fear	35	34	37	85	60.34	52.30	61.67

Emotion	Correct			Total	Recall		
	C1	C2a	C2b	C1	C1	C2a	C2b
sadness	46	45	41	267	17.22	16.85	15.36
joy	13	16	18	50	26	32	36.00
guilt	59	68	64	198	29.79	34.34	32.32
Total	225	251	264	895	25.13	28.04	29.50
Baseline	126	126	126	895	14.0.7	14.07	14.07

Table 6.9. Results of the emotion detection using EmotiNet on all test examples in test set C

Method	Mean Accuracy
DA: NB (with stemming)	67.2
DA: NB (without stemming)	67.4
DA: SVM (with stemming)	67.4
DA: SVM (without stemming)	66.9
DA SemEval: NB (with stemming)	F1= 27.9
DA SemEval: NB (without stemming)	F1= 28.5
DA SemEval: SVM (with stemming)	F1= 28.6
DA SemEval: SVM (without stemming)	F1= 27.8
DA SemEval: Vector Space Model (with stemming)	F1= 31.5
DA SemEval: Vector Space Model (without stemming)	F1= 32.2
EmotiNet A1	31.2
EmotiNet A2a	21.3
EmotiNet A2b	20.5
EmotiNet B1	26.16
EmotiNet B2a	41.56
EmotiNet B2b	40.05
EmotiNet C1	39.4
EmotiNet C2a	40.68
EmotiNet C2b	42.24
GENESIS	77.9

Table 6.10. Comparison of different systems for affect detection using ISEAR or self-reported affect in general

DISCUSSION AND CONCLUSIONS

From the results in Tables 6.4 to 6.9 we can conclude that the approach is valid, although much remains to be done to fully exploit the capabilities of EmotiNet. Given the number of core examples and the results obtained, we can see that the number of chains corresponding to one emotion in the core do influence the final result directly. However, the systems performs significantly better when an equal number of core examples is modeled, although when more emotions are evaluated (the difference between test sets B and C), the noise introduced leads to a drop in performance.

The comparative results shown in Table 6.10 show, on the one hand, that the task of detecting affect in text is very difficult. Thus, even if the appraisal criteria are directly given to a system (as in the case of GENESIS), its accuracy level only reaches up to 80%. If a system is trained on 90% of the data in one corpus using lexical information, its performance reaches up to around 68%. However, the results drop significantly when the approach is used on different data, showing that it is highly dependent on the vocabulary it uses. As opposed to this, the model we proposed based on appraisal theories proved to be flexible, its level of performance improving – either by percentual increase, or by the fact that the results for different emotional categories become more balanced. We showed that introducing new information can be easily done from existing common-sense knowledge bases and that the approach is robust in the face of the noise introduced.

From the error analysis we performed, we could determine some of the causes of error in the system. The first important finding is that extracting only the action, verb and patient semantic roles is not sufficient. There are other roles, such as the modifiers, which change the overall emotion in the text (e.g. "I had a fight with my sister" – sadness, versus "I had a fight with my stupid sister" – anger). Therefore, such modifiers should be included as attributes of the concepts identified in the roles, and, additionally, added to the tuples, as they can account for other appraisal criteria. This can also be a method to account for negation. Given that just 3 roles were extracted, there were also many examples that did not make sense when input into the system. Further on, we tried to assigned emotion to all the actions contained in the chains. However, some actions have no emotional effect. Therefore, an accurate source of knowledge on the affect associated to concepts has to be added.

Another issue we detected was that certain emotions tend to be classified most of the times as another emotion (e.g. fear is mostly classified as anger). This is due to the fact that emotions are subjective (one and the same situation can be a cause for anger for a person and a cause of fear to another or a mixture of the two); also, in certain situations, there are very subtle nuances that distinguish one emotion from the other.

A further source of errors was that lack of knowledge on specific actions. As we have seen, this knowledge can be imported from external knowledge bases and integrated in the core. This extension using larger common-sense knowledge bases, may lead to problems related to knowledge consistency and redundancy, with which we have not dealt with yet. VerbOcean extended the knowledge, in the sense that more examples could be classified. However, given the ambiguity of the resource and the fact that it is not perfectly accurate also introduced many errors.

Finally, other errors were produced by NLP processes and propagated at various steps of the processing chain (e.g. SRL, co-reference resolution). Some of these errors cannot be eliminated; however, others can be partially solved by using alternative NLP tools. A thorough analysis of the errors produced at each of the stages involved in the application and extension of EmotiNet must be made in order to obtain a clear idea of the importance/noise of each component.

6.4. CONCLUSIONS ON METHODS FOR IMPLICIT SENTIMENT DETECTION

In this second part of the chapter, we presented our contribution concerning three major topics. The first one was the proposal of a method to model real-life situations described in text based on the model of the appraisal theory. Based on the proposed model, the second contribution was the design and population of EmotiNet, a knowledge base of action chains representing and storing affective reaction to real-life contexts and situations described in text. We started our approach by modelling the actions presented in a set of phrases, describing affective situations, together with their affective value. We subsequently extended our model using VerbOcean. Finally, the third contribution lied in proposing and evaluating a method to detect emotion in text based on EmotiNet, using other examples in the ISEAR corpus.

We conclude that our approach is appropriate for detecting emotion in text, although additional elements should be included in the model and extra knowledge is required. Moreover, we found that the process of automatic evaluation was influenced by the low performance of the NLP tools used. Thus, alternative tools must be tested in order to improve the output. We must also test our approach on corpora where more than one emotion is assigned per context (such as the one in SemEval 2007 Task14).

All in all, we have seen that the issue of sentiment analysis cannot be resolved easily, by remaining at a word level or requiring the explicit mention of emotion or affect-related words. With this chapter, we completed our vision on the needs of sentiment analysis system and opened the door to a new direction for research, which is given by the necessity to develop methods that detect implicitly-expressed emotion.

CHAPTER 7. CONTRIBUTIONS

Motto: *"The value of the sentiment is the value of the sacrifice you are prepared to make for it." (John Galsworthy)*

This thesis focused on the resolution of different problems related to the task of sentiment analysis. Specifically, we concentrated on:

1. Defining the general task and related concepts, by presenting an overview of the present definition and clarifying the inconsistencies found among the ones that were previously given in the literature;
2. Proposing and evaluating methods to define and tackle sentiment analysis from a variety of textual genres, in different languages;
3. Redefining the task and proposing methods to annotate specific corpora for sentiment analysis in the corresponding text type, in different languages in case the task of sentiment analysis was not clearly defined for a specific textual genre and/or no specific corpora was available for it. These resources are publicly available for the use of the research community;
4. Applying opinion mining techniques in the context of end-to-end systems that involve other NLP tasks as well. To this aim, we concentrated on performing sentiment analysis in the context of question answering and summarization.
5. Carrying out experiments using existing question answering and summarization systems, designed to deal with factual data only.
6. Proposing and evaluating a new framework for what we called "opinion question answering" and new methods for "opinion summarization", subsequent to experiments showing that systems performing question answering and summarization over factual texts were not entirely suited in the context of opinions;
7. Presenting a general method for the detection of implicitly-expressed emotion from text. First, we presented the method to build a lexicon of terms that in themselves contain no emotion, but that trigger emotion in a reader. Subsequently, we abstracted from the analysis of sentiment expressed in text based on linguistic cues and proposed and evaluated a method to represent text as action chains. The emotion elicited by the situation presented in the text was subsequently judged using commonsense knowledge on the emotional effect of each action in the chain;
8. Evaluating our approaches in international competitions, in order to compare our approaches to others and validate them.

Further on, we present in detail the contributions we have made to the research in the field of sentiment analysis throughout this thesis and show how the methods and resources we proposed filled important gaps in the existing research. The main contributions answer five research questions:

1. *How can sentiment analysis and, in a broader perspective, opinion mining be defined in a correct way? What are the main concepts to be treated in order to do that?*

In Chapter 2, we presented a variety of definitions that were given to concepts related and involved in the task of sentiment analysis – subjectivity, objectivity, opinion, sentiment, emotion, attitude and appraisal. Our contribution in this chapter resided in clearly showing that sentiment analysis and opinion mining are not synonymous, although in the literature they are usually employed interchangeably. Additionally, we have shown that "opinion", as it is defined by the Webster dictionary, is not synonymous to sentiment. Whereas sentiments are types of opinions, reflecting the feelings (i.e. conscious part of emotions), all opinions are not sentiments (i.e there are types of opinions that are not reflective of emotions). We have also shown that subjectivity analysis is not directly linked to sentiment analysis as it is considered by many of the researchers in the field. In other words, detecting subjective sentences does not imply directly obtaining the sentences that contain sentiment. The latter, as expressions of evaluations based on emotion, are not necessarily indicated in subjective sentences, but can also be expressed in objective sentences. Subjective sentences can or cannot contain expressions of emotions. The idea is summarized in Chapter 2, Figure 2.1:

Finally, we have shown that there is a clear connection between the work done under the umbrella of sentiment analysis/opinion mining and the one in appraisal/attitude analysis. Although all these areas are usually considered to refer to the same type of work, the wider aim of attitude or appraisal analysis can capture much better the research that has been done in sentiment analysis, including all classes of evaluation (affective, cognitive, behavioral) and the connection between author, reader and text meaning. Based on this observation and in view of the Appraisal Theory, in Chapter 6 we proposed a model of emotion detection based on commonsense knowledge.

Clearly defining these concepts has also helped in defining in an appropriate manner the sentiment analysis task in the context of the different textual genres we employed in our research. Subsequently, the correct definition has made it possible to define annotation schemes and create resources for sentiment analysis in all the textual genres we performed research with. All these resources were consequently employed in the evaluation of the specific methods we created for sentiment analysis from different textual genres. Both the resources created, through the high inter-annotator agreement we obtained, as well as the methods we proposed, through the performance of the systems implementing them, have shown our efforts to give the clear definition were indeed an important contribution to this field.

2. *Can sentiment analysis be performed using the same methods, for all text types? What are the peculiarities of the different text types and how do they influence the methods to be used to tackle it? Do we need special resources for different text types?*
3. *Can the same language resources be used in other languages (through translation)? How can resources be extended to other languages?*

In Chapter 4, we showed the peculiarities of different text types (reviews, newspaper articles, blogs, political debates), analyzed them and proposed adequate techniques to address them at the time of performing sentiment analysis. We evaluated our approaches correspondingly and showed that they perform at the level of state-of-the-art systems and in many cases outperform them. In this chapter, we presented different methods and resources we built for the task of sentiment analysis in different text types. We started by presenting methods to tackle the task of feature-based opinion mining and summarization, applied to product reviews. We have analyzed the peculiarities of this task and identified the weak points of existing research. We proposed and evaluated different methods to overcome the identified pitfalls, among which the most important were the discovery of indirectly mentioned features and computing the polarity of opinions in a manner that is feature-dependent, using the Normalized Google Distance and Latent Semantic Analysis as measure of term association. Subsequently, we proposed a unified model for sentiment annotation for this type of text, able to capture the important phenomena that we had identified – different types of sentiment expressions – direct, indirect, implicit, feature mentioning and span of text expressing a specific opinion. Such a distinction had not been proposed in the literature. Its contribution is not only given by the annotation process, but also by the fact that considering larger spans of text as representing one single opinion has lead us to research on opinion question answering (see Chapter 5) using retrieval of 3-sentences-long text snippets, highly improving the performance of the opinion question answering system. Following this annotation scheme, we also proposed a method to detect and classify opinion stated on the most important features of a product, on which stars were given in the review, based on textual entailment. Apart from this contribution, the performance obtained applying this method is indicative of the fact that it is possible to obtain, apart from a 2-way classification of opinions on product features, a summary of the most important sentences referring to them. These latter can be employed to offer support snippets to the feature-based opinion mining and summarization process, which normally only offers a percent-based summary of the opinions expressed on the product in question.

Further on, we explored different methods to tackle sentiment analysis from newspaper articles. After the initial experiments, we analyzed the reasons for the low performance obtained and redefined the task, taking into account the peculiarities of this textual genre. We created an annotation model and labeled two different corpora of newspaper article quotations, in English and German. After redefining the task and delimiting the scope of the sentiment analysis process to quotations – small text snippets containing direct speech, whose source and target are previously known-, the annotation agreement rose significantly.

Additionally, improving the definition of the task made it possible to implement automatic processing methods that are appropriate for the task and significantly improve the performance of the sentiment analysis system we had designed. In the view of applying sentiment analysis to different types of texts, in which sentiment-bearing content is highly mixed with non-opinionated one and where the sources and targets of opinions are multiple, we have proposed different general methods for sentiment analysis, which we applied to political debates. The results of this latter experiment motivated us to analyze the requirements of a general labeling scheme for the task of sentiment analysis, which can be used to capture all relevant phenomena in sentiment expression. To this aim, in (Boldrini et al., 2009), we defined EmotiBlog, an annotation scheme that is able to capture, at a fine-grained level, all linguistic phenomena related to sentiment expression in text. The subsequent experiments have shown that this model is appropriate for the training of machine learning models for the task of sentiment analysis in different textual genres, in both languages in which experiments have been carried out using it – English and Spanish.

Finally, we have shown that the corpus annotated in this manner can be used to extract features for machine learning models that can be employed to tackle sentiment analysis in other languages, through translation. The good results obtained in the SemEval 2010 Task 18 – Disambiguation of Sentiment Ambiguous Adjectives competition, where we translated the texts from Traditional Chinese to English and applied an machine learning model trained on the EmotiBlog English data have proven that the annotation model is robust enough and useful even in the context of noisy data.

4. *How can we deal with opinion in the context of traditional tasks? How can we adapt traditional tasks (Information Retrieval, Question Answering, Text Summarization) in the context of opinionated content? What are the "new" challenges in this context?*

In Chapter 4, we have only concentrated on the task of sentiment analysis as a standalone challenge, omitting the steps required in order to obtain the texts on which the sentiment analysis methods were applied or eliminating redundancy in the information obtained. In a real-world application scenario, however, automatically detecting the opinion expressed in a text is often not the first, neither the last task to be performed. Prior to the analysis of the sentiment contained within, the relevant texts in which opinions are contained on required targets must be retrieved. Additionally, in many of the cases, the results obtained after automatically processing texts to determine the sentiment they contain still pose many problems in terms of volume. Thus, even if the sentiment is determined automatically, one may still require a summarization component, in order to further reduce the quantity of information, so that it is can be read and used by a person.

Bearing in mind these necessities, in Chapter 5 we researched on methods to combine opinion mining with question answering and summarization. Here, we have shown that performing traditional tasks in the context of opinionated text has many challenges and that systems that were designed to work exclusively with factual data are not able to cope with opinion questions. Our contribution resides in demonstrating, through evaluation, that in the case of such queries, for the question treatment, new elements have to be defined. We proposed the inclusion of the next elements: Expected Polarity Type, Expected Source, Expected Target, and have defined methods to detect them both from the question, as well as the candidate answers, using opinion mining techniques and employing Semantic Role Labeling. For the retrieval process, as we have shown before and confirmed in the OQA scenario, larger spans of texts are more appropriate when dealing with opinionated content. Specifically, we have shown that retrieving 3-sentences long snippets leads to better results in the case of OQA systems. By proposing and evaluating these new elements and techniques, we have shown the manner in which OQA can be tackled in a robust manner, consistent with the characteristics of opinionated texts. Lastly, we evaluated the impact of using different tools and resources in this task, for performing anaphora resolution and expanding the question through paraphrasing. Studying the manner in which these two components can be optimally added to the system is an important line for future work that we have opened.

In the case of opinion summarization, our contribution resides in: a) studying the order in which the opinion mining and the summarization systems have to be employed; b) studying the manner in which opinion mining and summarization systems can be used in tandem; c) studying the effect of topic detection in opinion mining in the context of opinion summarization; d) proposing a method to summarize opinion based on the intensity of the sentiment expressed.

In this part of the research, our contribution resides in showing that the sentiment analysis system must be employed prior to the summarization system and that the sentiment analysis component must be enhanced with topic-detection mechanisms. In other cases, although the sentiment analysis systems performs the classification correctly (according to the polarity of the sentiment), the fact that topic-relatedness is not contemplated leads to the introduction of irrelevant data in the final summaries. Finally, we have shown that in the case of opinionated text, relevance is given not only by the information contained, but also by the polarity of the opinion and its intensity. Although initial results have shown that there is no correlation between the Gold Standard annotations and the intensity level of sentences, as output by the sentiment analysis system, given the fact that using this method, we obtained high results as far as F-measure is concerned in TAC 2008, we believe that more mechanisms for opinion intensity should be studied, so that the clear connection between sentence relevance and the opinion it contains, as well as the intensity it has, can be established.

5. *Can we propose a model to detect emotion from text, in the cases where it is expressed implicitly, needing world knowledge?*

In the first chapters of this thesis, we explored the task of sentiment analysis in different text types and languages, proposing a variety of methods that were appropriate for tackling the issues in each particular text type. Most of the times, however, the approaches we took were limited to discovering only the situations where sentiment was expressed explicitly (i.e. where linguistic cues could be found in the text to indicate it contained subjective elements or sentiment). Nevertheless, in many cases, the emotion underlying the sentiment is not explicitly present in text, but is inferable based on commonsense knowledge (i.e. emotion is not explicitly, but implicitly expressed by the author, by presenting situations which most people, based on commonsense knowledge, associate with an emotion, like “going to a party”, “seeing your child taking his/her first step” etc.).

In Chapter 6 of the thesis, we presented our contribution to the issue of automatically detecting emotion expressed in text in an implicit manner. The initial approach is based on the idea that emotion is triggered by specific concepts, according to their *relevance,* seen in relation to the basic needs and motivations, underpinning our idea on the Relevance Theory. The second approach we propose is based on the Appraisal Theory models. The general idea behind it is that emotions are most of the times not explicitly stated in texts, but results from the interpretation (appraisal) of the actions contained in the situation described, as well as the properties of their actors and objects. Our contribution in this last part of the research resides in setting up a framework for representing situations described in text as chains of actions (with their corresponding actors and objects), and their corresponding properties (including the affective ones), as commonsense knowledge. We show the manner in which the so-called “appraisal criteria” can be automatically detected from text and how additional knowledge on the properties of the concepts involved in such situations can be imported from external sources.

Finally, we demonstrate through an extensive evaluation that such a representation is useful to obtain an accurate label of the emotion expressed in text, without any linguistic clue being present therein.

As sentiments are directly related to the presence of emotion, detecting implicit expressions of emotion can increase the performance of sentiment analysis systems, a fact that we have proven in the experiments we presented in Chapter 3 in the case of sentiment analysis in product reviews.

CHAPTER 8. FUTURE WORK

Motto: *"Nothing exists except atoms and empty space. Everything else is an opinion." (Democritus)*

In this final chapter, we present the possible directions for the further development of the research performed in this thesis. In order to be coherent with the structure of the thesis, these directions will be presented in relation to the research questions we aimed at answering and the points that remain to be addressed:

1. *How can sentiment analysis and, in a broader perspective, opinion mining be defined in a correct way? What are the main concepts to be treated in order to create a good definition that can be used to appropriately define the task and subsequently propose correct methods to tackle it?*

In Chapter 2 of the thesis, we presented an overview of the definitions given in the NLP literature to the related tasks of subjectivity analysis, sentiment analysis, opinion mining, appraisal/attitude analysis, emotion detection, as well as the concepts involved in these tasks. We subsequently showed that there is a high inconsistency between the definitions given both to the tasks, as well. Finally, we proposed an operational definition that was consistent with the manner in which the different terms related to sentiment analysis were defined in well-established sources and in a manner that was coherent with the approaches we took in the research. Subsequently, in Chapter 4, we showed that sentiment analysis must be tackled in a different manner, depending on the types of text considered. From these efforts, a future line of work is the definition of a unified framework for sentiment analysis, that describes the task in a general, yet consistent manner across genres and applications. In this sense, as we have shown in the case of newspaper articles, such a framework should be built by taking into account not only the textual content, but also the elements that relate to the author and reader.

The subsequent research question we addressed in this thesis were:

2. *Can sentiment analysis be performed using the same methods, for all text types? What are the peculiarities of the different text types and how do they influence the methods to be used to tackle it? Do we need special resources for different text types?*
3. *Can the same language resources be used in other languages (through translation)? How can resources be extended to other languages?*

In Chapter 4, we presented different methods and resources for the task of sentiment analysis in different text types (reviews, newspaper articles, blogs, political debates), in different languages (English, Spanish, German). A future line of work is the extension of the proposed resources for other languages, either through translation, in which case the resources should be refined and evaluated on different types of texts in their original language, or by direct annotation in the target language. In the latter case, it would be interesting to study the differences given by the peculiarities of sentiment expression in a manner that is dependent on the culture and language differences. In direct relation to the work developed in this thesis, in the different text types and languages, future lines of research could be:

1. In the case of review texts:
    a. The automatic extraction of taxonomies for product features
    b. The extension of the proposed framework for review annotation and feature-based opinion mining and summarization for languages other than English and Spanish
2. In the context of newspaper quotations and, in a more general manner, newspaper articles:
    a. The study of the impact of news source (i.e. in terms of bias, reputation, trust) on the sentiment analysis process, within sources from the same country/culture and across countries and cultures
    b. The study of the influence that the reader background has on the manner in which sentiment is perceived from newspaper articles
    c. The automatic and semi-automatic extension of the resources we built to other languages (i.e. in addition to the collection of quotations we annotated for English and German)
    d. The development of a framework for sentiment analysis that takes into account the 3 proposed components of text – the author, reader and text – and the manner in which they interact in the text meaning negotiation, according to the Speech-Act and the Appraisal Theories
    e. The study of the influence of news content (i.e. what we denoted as "good versus bad news") on the manner in which sentiment is expressed and subsequently on the performance of the automatic sentiment analysis task
3. In the context of political debates and general texts:
    a. Study the methods to represent the dialogue structure and the influence of different methods for discourse analysis, as well as tools (e.g. for anaphora resolution) on the proposed sentiment analysis methods.
    b. The study and use of topic modeling techniques in order to accurately detect the target of the sentiment expressed, independent of its nature being known beforehand

    c. The study of affect-based argumentation techniques in order to detect the use of emotion triggers in relation to the discussion topic

4. In the context of blogs:
    a. The extension of the proposed method of annotation and labeled corpora to other languages and similar text types (e.g. forums, discussion boards)
    b. The study and use of topic modeling techniques, as well as of the impact of co-reference resolution at an inter-textual level

4. *How can we deal with opinion in the context of traditional tasks? How can we adapt traditional tasks (Information Retrieval, Question Answering, Text Summarization) in the context of opinionated content? What are the "new" challenges in this context?*

Bearing in mind these necessities of real-world applications, which need, apart from the sentiment analysis component, also text retrieval and text summarization components, in Chapter 5 we proposed methods to combine opinion mining with question answering and summarization. We showed that performing traditional tasks in the context of opinionated text has many challenges and that systems that were designed to work exclusively with factual data are not able to cope with opinion questions. In this thesis, we proposed new methods and techniques to adapt question answering and summarization systems to deal with opinionated content.

In the case of opinion question answering systems, future work includes the development of a benchmark for opinion questions' classification and the proposal of adequate methods to tackle each type of opinion queries, in a monolingual, multilingual and cross-lingual setting. Additionally, the framework for opinion question answering should be extended with appropriate resources to other languages. Further on, as we have seen from our experiments in the NTCIR 8 MOAT competition, there is an immediate need to include high-performing methods for temporal expression resolution and anaphora resolution. Unfortunately, due to the low performance of systems resolving these aspects, at this point the influence they have on the opinion question answering system's performance is negative. Another line of future work is the study of query expansion techniques that are appropriate for opinionated content. From what we have seen in our experiments, the use of a paraphrase collection that is not specifically designed for the sentiment-bearing textual content leads to a drop in performance of the final system.

In the case of sentiment summarization approaches, the lines for future work include the development of appropriate topic-sentiment relevance detection techniques and the study of qualitative measures for the evaluation of the opinion summarization approaches.

The last research question we addressed in this thesis was:

5. *Can we propose a model to detect emotion (as a component of sentiment) from text, in the cases where it is expressed implicitly, requiring world knowledge for its detection?*

In Chapter 6 of the thesis, we proposed methods to detect implicit expressions of emotion, from situations where the affective label is only inferable based on commonsense knowledge (i.e. emotion is not explicitly, but implicitly expressed by the author, by presenting situations which most people, based on commonsense knowledge, their needs and motivations, associate with an emotion, like "war", "terrorism", "going to a party", "seeing your child taking his/her first step" etc.). The initial approach we proposed is based on the idea that emotion is triggered by specific concepts, according to their *relevance,* seen in relation to the basic needs and motivations. Subsequently, we propose a framework for affect detection in text based on the Appraisal Theory, modeling the possible interpretation (appraisal) of the actions contained in the situation described, according to the commonsense knowledge on the actions and the properties of their actors and objects. Future work includes the extension of the created resources to other languages and the expansion of the common-sense knowledge base of emotion eliciting situation with additional external sources of knowledge. In this context, interesting lines for future work could be:

- The development techniques for the personalization and adaptation of the information content depending on the user context, through the exploitation of the subjective content generated by the user, as well as the analysis of the preferences expressed explicitly, through options in user profiles and implicitly, through the connections in social networks, contributions to forums, reviews, blogs, microblogs.
- The design, implementation, population and extension of a knowledge base for user preference modeling, based on which opinion mining can be performed in a personalized manner (i.e. depending on the user's goals, beliefs, opinions, views, feelings). In this context, the semantic search capabilities can be improved and extended in the context of subjective information, in a user-centric manner.

REFERENCES

1. [Agichtein and Gravano, 2000] Agichtein, D. and Gravano, L. (2000). Snowball: Extracting Relations from Large Plain-Text Collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries (ACM DL)* 2000: 85-94.
2. [Atkinson and Van der Goot, 2009] Atkinson, M. and Van der Goot, E. (2009). Near real time information mining in multilingual news. In *Proceedings of the 18th International World Wide Web Conference***,** 2009: 1153-1154.
3. [Austin, 1976] Austin, J.L. (1976). *How to do things with words.* Oxford et. al.: Oxford University Press.
4. [Balahur and Montoyo, 2008] Balahur, A. and Montoyo, A. Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification. In *Proceedings of the AISB 2008 Convention "Communication, Interaction and Social Intelligence", 2008.*
5. [Balahur and Montoyo, 2008a] Balahur, A. and Montoyo, A. (2008a). Applying a culture dependent emotion triggers database for text valence and emotion classification. In *Procesamiento del Lenguaje Natural, 40(40).*
6. [Balahur and Montoyo, 2008b] Balahur, A. and Montoyo, A. (2008b). Building a recommender system using community level social filtering. In *5th International Workshop on Natural Language and Cognitive Science (NLPCS),* 2008:32-41.
7. [Balahur and Montoyo, 2008c] Balahur, A. and Montoyo, A. (2008c). Determining the semantic orientation of opinions on products - a comparative analysis. In *Procesamiento del Lenguaje Natural***,** 41(41).
8. [Balahur and Montoyo, 2008d] Balahur, A. and Montoyo, A. (2008d). A feature-driven approach to opinion mining and classification. In *Proceedings of the NLPKE 2008.*
9. [Balahur and Montoyo, 2008e] Balahur, A. and Montoyo, A. (2008e). An incremental multilingual approach to forming a culture dependent emotion triggers lexical database. In *Proceedings of the Conference of Terminology and Knowledge Engineering (TKE 2008).*
10. [Balahur and Montoyo, 2008f] Balahur, A. and Montoyo, A. (2008f). Multilingual feature-driven opinion extraction and summarization from customer reviews .In *Lecture Notes in Computer Science, 5039, NLDB,* 2008:345-346.
11. [Balahur and Montoyo, 2009] Balahur, A. and Montoyo, A. (2009). Semantic approaches to fine and coarse-grained feature-based opinion

mining. In *Proceedings of the International Conference on Application of Natural Language to Information Systems , NLDB* 2009:142-153

12. [Balahur and Montoyo, 2010] Balahur, A.; Montoyo, A. OpAL: Applying Opinion Mining Techniques for the Disambiguation of Sentiment Ambiguous Adjectives in SemEval-2 Task 18. In *Proceedings of SemEval-2, the 5th International Workshop on Semantic Evaluation, satellite workshop to ACL 2010.*
13. [Balahur and Steinberger, 2009] Balahur, A. and Steinberger, R. (2009). Rethinking Opinion Mining in Newspaper Articles: from Theory to Practice and Back. In *Proceedings of the first workshop on Opinion Mining and Sentiment Analysis (WOMSA 2009).*
14. [Balahur et al, 2010] Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. (2010). *OpAL: a System for Mining Opinion from Text for Business Applications.* In Marta E. Zorrilla, Jose-Norberto Mazón, Óscar Ferrández, Irene Garrigós, Florian Daniel and Juan Trujillo (Eds.). *Business Intelligence Applications and the Web: Models, Systems and Technologies*, IGI Global Press.
15. [Balahur et al., 2008] Balahur, A., Lloret, E., Ferrández, O., Montoyo, A., Palomar, M., and Muñoz, R. (2008). The DLSIUAES Team's Participation in the TAC 2008 Tracks. In *Proceedings of the Text Analysis Conference* .
16. [Balahur et al., 2009a] Balahur, A., Boldrini, E., Montoyo, A., and Martínez-Barco, P. (2009). A comparative study of open domain and opinion question answering systems for factual and opinionated queries. In *Proceedings of the Conference on Recent Advances in Natural Language Processing 2009*.
17. [Balahur et al., 2009b] Balahur, A., Boldrini, E., Montoyo, A., and Martínez-Barco, P. (2009). Cross-topic opinion mining for real-time human-computer interaction. In *Proceedings of the Workshop on Natural Language and Cognitive Science , NLPCS 2009*: 13-22.
18. [Balahur et al., 2009c] Balahur, A., Boldrini, E., Montoyo, A., and Martínez-Barco, P. (2009). Fact versus opinion question classification and answering: Challenges and keys. In *Proceedings of the International Conference on Artificial Intelligence , ICAI 2009:* 750-755.
19. [Balahur et al., 2009d] Balahur, A., Boldrini, E., Montoyo, A., and Martínez-Barco, P. (2009). Opinion and generic question answering systems: a performance analysis. In *Proceedings of the ACL-IJCNLP* 2009: 157-160.
20. [Balahur et al., 2009e] Balahur, A., Kozareva, Z., and Montoyo, A. (2009). Determining the polarity and source of opinions expressed in

political debates. *Lecture Notes in Computer Science, 5449,CICLing* 2009:468-480.

21. [Balahur et al., 2009f] Balahur, A., Steinberger, R., Van der Goot, E., and Pouliquen, B. (2009). Opinion mining from newspaper quotations. In *Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content*, IAT Workshops,2009: 523-526.
22. [Balahur et al., 2009g] Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R., Montoyo, A. (2009). Summarizing Opinions in Blog Threads. *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC), pp. 606-613, Hong Kong.*
23. [Balahur et al., 2009h] Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. (2009). Evaluating Question Answering Systems on Opinion Queries. *Proceedings of TSA 2009, 1st CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement.*
24. [Balahur et al., 2009i] Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., Martínez-Barco, P. (2009). Summarizing Threads in Blogs Using Opinion Polarity. *Proceedings of the EETTs 2009 Workshop, satellite to the RANLP 2009 Conference.*
25. [Balahur et al., 2010a] Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. (2010). Going Beyond Traditional QA Systems: Challenges and Keys in Opinion Question Answering. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010).*
26. [Balahur et al., 2010b] Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. (2010). The OpAL System at NTCIR 8 MOAT. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR 8).*
27. [Balahur et al., 2010c] Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. (2010). Opinion Question Answering: Towards a Unified Approach. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010), 511-516.*
28. [Balahur et al., 2010d] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J. (2010). Sentiment Analysis in the News. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta.*
29. [Balahur et al., 2010e] Balahur, A., Kabadjov, M., Steinberger, J. (2010). Exploiting Higher-level Semantic Information for the Opinion-oriented Summarization of Blogs. *International Journal of*

Computational Linguistics and Applications, ISSN: 0976-0962, Vol. 1, No. 1-2, pp. 45-59.

30.[Balahur et al., 2011a] Balahur, A., Hermida, J.M., Montoyo, A. (2011). Detecting Emotions in Social Affective Situations Using the EmotiNet Knowledge Base. In *Lecture Notes in Computer Science Nr. 6677, proceedings of the 8th International Symposium on Neural Networks ISNN 2011*.

31.[Balahur et al., 2011b] Balahur, A., Hermida, J.M., Montoyo, A. (2011). EmotiNet: a Knowledge Base for Emotion Detection in Text Build on the Appraisal Theories. In *Lecture Notes in Computer Science (in press), proceedings of NLDB 2011*.

32.[Banea et al., 2008a] Banea, C., Mihalcea, R., and Wiebe, J. (2008a). A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Conference on Language Resources and Evaluations (LREC 2008)*, Maraakesh, Marocco.

33.[Banea et al., 2008b] Banea, C., Mihalcea, R., Wiebe, J., and Hassan, S. (2008b). Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, 127-135, Honolulu,Hawaii.

34.[Banea et al., 2010] Banea, C., Mihalcea, R. and Wiebe, J. (2010). Multilingual subjectivity: are more languages better? In *Proceedings of the International Conference on Computational Linguistics (COLING 2010)*, p. 28-36, Beijing, China,

35.[Banfield, 1982] Banfield, A. (1982). *Unspeakable sentences: Narration and Representation in the Language of Fiction*. Routledge and Kegan Paul.

36.[Bansal et al., 2008] Bansal, M., Cardie, C. and Lee, L. (2008). The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. In *Proceedings of the International Conference on Computational Linguistics (COLING), 2008. Poster paper*, pp.15-18.

37.[Berland and Charniak, 1999] Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of ACL, 1999*, College Park, Md, pp.57-64.

38.[Boldrini et al., 2009] Boldrini, E., Balahur, A., Martínez-Barco, P., and Montoyo, (2009). Emotiblog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In *Proceedings of the 5th International Conference on Data Mining (DMIN 2009)*, 491-497.

39.[Boldrini et al., 2010] Boldrini, E., Balahur, A., Martínez-Barco, P., Montoyo, A. (2010). EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV)*, satellite workshop to ACL 2010, pp. 1-10, Uppsala Sweden.

40.[Bossard et al., 2008] Bossard, A., Généreux, M., and Poibeau, T. (2008). Description of the LIPN systems at TAC 2008: Summarizing information and opinions. In *Proceedings of the Text Analysis Conference of the National Institute for Standards and Technology*, Geithersbury, Maryland, USA.

41.[Breckler and Wiggins, 1992] Breckler, S. J. and Wiggins, E. C. (1992). On defining attitude and attitude theory: Once more with feeling. In A. R. Pratkanis, S. J. Breckler, and A. C. Greenwald (Eds.), *Attitude structure and function*. Hillsdale, NJ, Erlbaum. pp. 407–427.

42.[Brin, 1998] Brin, S. (1998). Extracting patterns and relations from the World-Wide Web. In *Proceedings of the 1998 International Workshop on Web and Databases (WebDB'98),* 1998: 172-183.

43.[Calvo and D'Mello, 2010] Calvo, R. A. and D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods and Their Applications. *IEEE Transactions on Affective Computing*, Vol. 1, No. 1, Jan.-Jun 2010, pp.18-37.

44.[Cambria et al., 2009] Cambria, E., Hussain, A., Havasi, C. and Eckl, C. (2009). Affective Space: Blending Common Sense and Affective Knowledge to Perform Emotive Reasoning. *In Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA)* 2009: 32-41.

45.[Cardie et al., 2004] Cardie, C., Wiebe, J., Wilson, T., Litman, D. (2004). Low-Level Annotations and Summary Representations of Opinions for Multiperspective QA. In *Mark Maybury (ed), New Directions in Question Answering* , AAAI Press/MIT Press, pp.17-98.

46.[Cerini et al., 2007] Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. (2007). Micro-wnop: A gold standard for the evaluation of auto-matically compiled lexical resources for opinion mining, Milano, IT.

47.[Chaovalit and Zhou, 2005] Chaovalit, P. and Zhou, L. (2005). Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. In *Proceedings of HICSS-05, the 38th Hawaii International Conference on System Sciences*, IEEE Computer Society.

48.[Chklovski and Pantel, 2004] Chklovski, T. and Pantel, P. (2004). VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of EMNLP, 2004*: 33-40.

49.[Choi et al., 2005] Choi, Y., Cardie, C., Rilloff, E., Padwardhan, S. (2005). Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *Proceedings of the HLT/EMNLP*.

50.[Cilibrasi and Vitanyi, 2006a] Cilibrasi, D. and Vitanyi, P. (2006a). Automatic Meaning Discovery Using Google. *IEEE Journal of Transactions on Knowledge and Data Engineering*.

51.[Conroy and Schlesinger, 2008] Conroy, J. and Schlesinger, S. (2008). Classy at TAC 2008 metrics. In *Proceedings of the Text Analysis Conference of the National Institute for Standards and Technology*.

52.[Cruz et al., 2008] Cruz, F., Troyano, J., Ortega, J., and Enríquez, F. (2008). The Italica system at TAC 2008 opinion summarization task. In *Proceedings of the Text Analysis Conference of the National Institute for Standards and Technology*.

53.[Cui et al., 2006] Cui, H., Mittal, V., and Datar, M. (2006). Comparative Experiments on Sentiment Classification for Online Product Reviews. In *Proceedings of the 21st National Conference on Artificial Intelligence AAAI 2006*, pp.1265-1270.

54.[Cytowic, 1989] Cytowic, R. (1989). *Synesthesia. A Union of the Senses*, New York, Springer Verlag.

55.[Dagan et al., 2006] Dagan, I., Glickman, O. and Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In *Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.) Machine Learning Challenges. Lecture Notes in Computer Science* , Vol. 3944, pp. 177-190, Springer.

56.[Danisman and Alpkocak, 2008] Danisman, T. and Alpkocak, A. (2008). Feeler: Emotion Classification of Text Using Vector Space Model. In *Proceedings of the AISB 2008 Convention, "Communication, Interaction and Social Intelligence*", vol.2, pp.53-59, Aberdeen,UK.

57.[Das and Chen, 2001] Das, S. and Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).*

58.[Das and Chen, 2007] Das, S.R. and Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the Web. *Management Science, 53(9)*, pp.1375–1388.

59.[Dave et al., 2003] Dave, K., Lawrence, S., and Pennock, D. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic

Classification of Product Reviews. In *Proceedings of WWW-2003*, 519-528.

60. [De Rivera, 1977] De Rivera, J. (1977). A structural theory of the emotions. *Psychological Issues*, 10 (4), Monograph 40, New York, International Universities Press.
61. [Deerwester et al., 1990] Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science,* 3(41).
62. [Denecke, 2008] Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. In *Proceedings of the International Conference on Data Engineering (ICDE 2008), Workshop on Data Engineering for Blogs, Social Media, and Web 2.0.,*2008: 430-435.
63. [Devitt and Ahmad, 2007] Devitt, A. and Ahmad, K. (2007). A lexicon for polarity: Affective content in financial news text. In *Proceedings of the Conference on Language for Special Purposes*, ACL 2007.
64. [Ding et al., 2008] Ding, X., Liu, B., and Yu, P. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of WSDM 2008*, New York, USA.
65. [Dyer, 1987] Dyer, M. (1987). Emotions and their computations: three computer models. *Cognition and Emotion*, 1, 323-347.
66. [Edmonds and Hirst, 2002] Edmonds, P. and Graeme, H. (2002). Near-synonymy and lexical choice. *Computational Linguistics, 28(2), June 2002, 105—144.*
67. [Ekman, 1999] Ekman, P. (1999). Basic Emotions. In *T. Dalgleish and M. Power (Eds.). Handbook of Cognition and Emotion.* Sussex, U.K.: John Wiley & Sons, Ltd., 1999.
68. [Erkan and Radev, 2004] Erkan, G. and Radev, D. R. (2004). LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 22,457-479.
69. [Esuli and Sebastiani, 2005] Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss analysis. In *Proceedings of CIKM* 2005: 617-624, Bremen, Germany.
70. [Esuli and Sebastiani, 2006] Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available resource for opinion mining. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp.417-422.
71. [Evans, 2001] Evans, D. (2001). Emotions. The science of sentiment. Oxford: Oxford University Press.

72.[Fellbaum, 1999] Fellbaum, C. (1999). WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press.

73.[Ferrández et al., 1999] Ferrández, A., Palomar, M., and Moreno, L. (1999). An Empirical Approach to Spanish Anaphora Resolution. *Machine Translation. Special Issue on Anaphora Resolution* In Machine Translation, pages 191–216.

74.[Frijda, 1986] Frijda, N. (1986). The emotions. Cambridge University Press.

75.[Gamon et al., 2005] Gamon, M., Aue, S., Corston-Oliver, S., and Ringger, E. (2005). Mining Customer Opinions from Free Text. *Lecture Notes in Computer Science, IDA* 2005:121-132.

76.[Gamon, 2004] Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING) 2004*, pp.841-847.

77.[Goldberg and Zhu, 2006] Goldberg, A. B. and Zhu, J. (2006). Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing.*

78.[Goldie, 2000] Goldie, P. (2000) *Emotions. A Philosophical Exploration*. Oxford University Press.

79. [Goleman, 1995] Goleman, D. (1995*). Emotional Intelligence*. Bantam Books, New York, USA.

80.[Gong and Liu, 2002] Gong, Y. and Liu, X. (2002). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of ACM SIGIR*, 2002: 19-25, New Orleans, USA.

81.[Grüninger and Fox, 1995] Grüninger, M. and Fox, M. S. (1995). Formal ontology in information systems. In *Proceedings of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal.

82.[Halliday, 1994] Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*. Edward Arnold, London,UK.

83.[Hatzivassiloglou and McKeown, 1997] Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, ACL 1997: 174-181, Madrid, Spain.

84.[Hatzivassiloglou and Wiebe, 2000] Hatzivassiloglou, V. and Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. *In Proceedings of COLING* 2000: 299-305..

85. [He et al., 2008] He, T., Chen, J., Gui, Z., and Li, F. (2008). CCNU at TAC 2008: Proceeding on using semantic method for automated summarization yield. In *Proceedings of the Text Analysis Conference of the National Institute for Standards and Technology*.
86. [Hovy, 2005] Hovy, E. H. (2005). Automated text summarization. In *Mitkov, R., editor, The Oxford Handbook of Computational Linguistics,* pages 583–598. Oxford University Press, Oxford, UK.
87. [Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining Opinion Features in Customer Reviews. In *Proceedings of Nineteenth National Conference on Artificial Intellgience AAAI-2004,* San Jose, USA.
88. [Iftene and Balahur-Dobrescu, 2007] Iftene, A. and Balahur-Dobrescu, A. (2007). Hypothesis transformation and semantic variability rules used in recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing 2007*.
89. [Jiang and Vidal, 2006] Jiang, H. and Vidal, J. M. (2006). From rational to emotional agents. In *Proceedings of the AAAI Workshop on Cognitive Modeling and Agent-based Social Simulation.*
90. [Jijkoun et al., 2010] Jijkoun, V., de Rijke, M., Weerkamp, W. (2010). Generating Focused Topic-Specific Sentiment Lexicons. ACL 2010, pp. 585-594.
91. [Johnson-Laird and Oatley, 1989] Johnson-Laird, P. N. and Oatley, K. (1989). The language of emotions: An analysis of a semantic field. Cognition and Emotion, 3, 81-123.
92. [Kabadjov et al, 2009a] Kabadjov, M. A., Steinberger, J., Pouliquen, B., Steinberger, R., and Poesio, M. (2009). Multilingual statistical news summarisation: Preliminary experimentswith english. In *Proceedings of theWorkshop on Intelligent Analysis and Processing of Web News Content at the IEEE / WIC / ACM International Conferenceson Web Intelligence and Intelligent Agent Technology(WI-IAT),* 2009:519-522.
93. [Kabajov et al., 2009] Kabadjov, M., Balahur, A., Boldrini, E. (2009). Sentiment Intensity: Is It a Good Summary Indicator? *Proceedings of the 4th Language Technology Conference LTC,* 2009: 203-212.
94. [Kim and Hovy, 2004] Kim, S.-M. and Hovy, E. (2004). Determining the Sentiment of Opinions. In *Proceedings of COLING 2004*, pp.1367-1373, Geneva, Swizerland.
95. [Kim and Hovy, 2005] Kim, S.-M. and Hovy, E. (2005). Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP),* Jeju Island, Korea.

96. [Kim and Hovy, 2006] Kim, S.-M. and Hovy, E. (2006). Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL Main Conference Poster Sessions, pp. 483–490.*
97. *[Kim et al., 2010] Kim, J., Li, J.-J. and Lee, J.-H. (2010). Evaluating Multilanguage-Comparability of Subjectivity Analysis Systems. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 595–603, Uppsala, Sweden, 11-16 July 2010.*
98. [Kintsch, 1999] Kintsch, W. (1999). *Comprehension: A paradigm for cognition*. New York, Cambridge University Press.
99. [Kintsch, 2000] Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review*, 7, 527-566.
100. [Koppel and Shtrimberg, 2004] Koppel, M. and Shtrimberg, I. (2004). Good or bad news? Let the market decide. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applicationns.*
101. *[Kouleykov and Magnini, 2006] Kouleykov, M. and Magnini, B. (2006). Tree Edit Distance for Recognizing Textual Entailment: Estimating the Cost of Insertion. In* Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.*
102. *[Kozareva et al., 2007] Kozareva, Z., Vázquez, S., and Montoyo, A. (2007). Discovering the Underlying Meanings and Categories of a Name through Domain and Semantic Information. In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP 2007), Borovetz, Bulgaria.*
103. [Ku et al., 2005] Ku, L.-W., Li, L.-Y., Wu, T.-H., and Chen, H.-H. (2005). Major topic detection and its application to opinion summarization. In *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR),* Salvador, Brasil.
104. [Ku et al., 2006] Ku, L.-W., Liang, Y.-T., and Chen, H.-H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI CAAW 2006),* AAAI Technical Report, SS-06-03, Standford University, 100-107.
105. [Kudo and Matsumoto, 2004] Kudo, T. and Matsumoto, Y. (2004). A boosting algorithm for classification of semistructured text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2004.*

106. [Lappin and Leass, 1994] Lappin, S. and Leass, H.J. (1994). An algorithm for pronominal anaphora resolution. *In Journal of Computational Linguistics*, 20(4):535–561.

107. [Larousse, 1995] Larousse (1995). *Diccionario Ideológico de la Lengua Española*. Larousse Editorial, RBA Promociones Editoriales, S.L.

108. [Laver et al., 2003] Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review* , 97(2, May): 311-331.

109. [Lazarus and Smith, 1988] Lazarus, R. S. and Smith, C. A. (1988). Knowledge and appraisal in the cognition-emotion relationship. *Cognition and Emotion*, 2, 281-300.

110. [Lee, 2004] Lee, L. (2004). "I'm sorry Dave, I'm afraid I can't do that": Linguistics, Statistics, and Natural Language Processing circa 2001. In *Computer Science: Reflections on the Field, Reflections from the Field (Report of the National Academies' Study on the Fundamentals of Computer Science),* pp. 111-118, 2004.

111. [Li et al., 2008] Li, W., Ouyang, Y., Hu, Y., Wei, F. (2008). PolyU at TAC 2008. In Proceedings of the Text Analysis Conference 2008, National Institute for Science and Technology, Gaithersburg, Maryland, USA .

112. [Li et al., 2008a] Li, F., Zheng, Z., Yang, T., Bu, F., Ge, R., Yan Zhu, X., Zhang, X., and Huang, M. (2008a). THU QUANTA at TAC 2008 QA and RTE track. In *Proceedings of the Text Analysis Conference (TAC 2008).*

113. [Li et al., 2008b] Li, W., Ouyang, Y., Hu, Y., and Wei, F. (2008b). PolyU at TAC 2008. In *Proceedings of the Text Analysis Conference 2008.*

114. [Lin and Pantel, 2001] Lin, D. and Pantel, P. (2001). Discovery of Inference Rules for Question Answering. In Journal of Natural Language Engineering 7(4):343-360.

115. [Lin et al., 2006] Lin, W., Wilson, T., Wiebe, J., and Hauptman, A. (2006). Which Side are You on? Identifying Perspectives at the Document and Sentence Levels. In *Proceedings of the Tenth Conference on Natural Language Learning CoNLL'06.*

116. [Lin, 1998] Lin, D. (1998). Dependency-based Evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems,*Granada, Spain, May,1998.

117. [Lita et al., 2005] Lita, L., Schlaikjer, A., Hong, W., and Nyberg, E. (2005). Qualitative dimensions in question answering: Extending the definitional QA task. In Proceedings of AAAI 2005.

118. [Liu and Maes, 2007] Liu, H. and Maes, P. (2007). Introduction to the semantics of people and culture, *International Journal on Semantic Web and Information Systems, Special Issue on Semantics of People and Culture (Eds. H. Liu & P. Maes), 3(1),* Hersey, PA: Idea Publishing Group.

119. [Liu and Singh, 2004] Liu, H. and Singh, P. (2004). ConceptNet: A Practical Commonsense Reasoning Toolkit. BT Technology Journal, Volume 22, nr.4, pp.211-226, Kluwer Academic Publishers.

120. [Liu et al., 2003] Liu, H., Lieberman, H. and Selker, T. (2003). A Model of Textual Affect Sensing Using Real-World Knowledge. *In Proceedings of IUI 2003.*

121. [Liu, 2007] Liu, B. (2007). *Web Data Mining. Exploring Hyperlinks, Contents and Usage Data*. Springer, first edition.

122. [Liu, 2010] Liu, B. (2010). Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing,* eds. N.Indurkhya and F.J.Damenan, 2010.

123. [Lloret et al., 2008] Lloret, E., Ferrández, O., Muñoz, R., and Palomar, M. (2008). A text summarization approach under the influence of textual entailment. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2008*).

124. [Lloret et al., 2009] Lloret, E., Balahur, A., Palomar, M., and Montoyo, A. (2009). Towards building a competitive opinion summarization system: Challenges and keys. In *Proceedings of the NAACL-HLT 2009 Student Research Workshop*, 2009: 72-77.

125. [Macleod et al., 1994] Macleod, C., Grishman, R., and Meyers, A. (1994). Creating a common syntactic dictionary of english. In *Proceedings of the International Workshop on Sharable Natural Language Resources*.

126. [Macleod et al., 1998] Macleod, C., Grishman, R., Meyers, A., Barrett, L., and Reeves, R. (1998). Nomlex: A lexicon of nominalizations. In *Proceedings of EURALEX 1998,* Liege, Belgium.

127. [Martin and Vanberg, 2007] Martin, L. W. and Vanberg,G. (2007). Coalition Government and Political Communication. In Political Research Quarterly, September 2008, Vol. 61, No. 3, pp. 502-516.

128. [Martin and Vanberg, 2008] Martin, L. W. and Vanberg, G. (2008). A robust transformation procedure for interpreting political text. Political Analysis Advance Access, Oxford University Press, 16(1).

129. [Martin and White, 2005] Martin, J. R. and White, P. R. (2005). Language of Evaluation: Appraisal in English. Palgrave Macmillan.

130. [Martín et al., 2010] Martín, T., Balahur, A., Montoyo, A., Pons, A. (2010). Word Sense Disambiguation in Opinion Mining: Pros and Cons. In *Journal Research in Computing Science, Special Issue: Natural Language Processing and its Applications,* ISSN: 1870-4069, Vol. 46, pp. 119-130.

131. [Martín-Wanton et al., 2010] Martín-Wanton, T., Pons-Porrata, A., Montoyo-Guijarro, A., Balahur, A. (2010). Opinion Polarity Detection – Using Word Sense Disambiguation to Determine the Polarity of Opinions. In *Proceedings of the 2ndInternational Conference on Agents and Artificial Intelligence (ICAART 2010).*

132. [Maslow, 1943] Maslow, A. (1943). A theory of human motivation. *Psychological Review*, 1943, Vol. 50, pp. 370–396.

133. [Masum Shaikh et al., 2007] Masum Shaikh, A., Prendinger, M., and Mitsuru, (2007). Assessing sentiment of text by semantic dependency and contextual valence analysis. *Lecture Notes in Computer Science Nr. 4738, ACII,* 2007: 737-738.

134. [Matsumoto et al., 2005] Matsumoto, S., Takamura, H., and Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD*, 2005: 301-311.

135. [Matthiassen, 1995] Matthiassen, C. (1995). Lexico-grammatical cartography: English systems. International Language Science Publishers.

136. [Max-Neef, 1991] Max-Neef, M. A. (1991). Human scale development: conception, application and further reflections. The Apex Press, New York.

137. [McCarthy, 1959] McCarthy, J. Programs with Common Sense. In Mechanisation of Thought Processes, *Proceedings of the Symposium of the National Physics Laborator*y, pages 77-84, London, U.K., 1959. Her Majesty's Stationery Office. Reprinted in John McCarthy. Formalizing Common Sense: Papers by John McCarthy. Ablex Publishing Corporation, 1990.

138. [McDonald et al., 2007] McDonald, R., Hannan, K., Neylon, T., Wells, M., Reynar, J. C. (2007). Structured Models for Fine-to-Coarse Sentiment Analysis. In *Proceedings of ACL 2007.*

139. [Mei Lee et al., 2009] Mei Lee, S. Y., Chen, Y. and Huang, C.-R. (2009). Cause Event Representations of Happiness and Surprise. In *Proceedings of PACLIC 2009,* Hong Kong.

140. [Mihalcea et al., 2007] Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the Conference of the Annual Meeting of the Association for Computational Linguistics 2007,* pp.976-983, Prague, Czech Republic.

141. [Moreda et al., 2007] Moreda, P., Navarro, B. and Palomar, M. (2007). Corpus-based semantic role approach in information retrieval", *Data Knowledge English. (DKE)* 61(3): 467-483.

142. [Moreda et al., 2008a] Moreda, P., Llorens, H., Saquete, E., and Palomar, M. (2008a). Automatic generalization of a qa answer extraction module based on semantic roles. In *Proceedings of AAI - IBERAMIA* , vol. 5290, 2008: 233-242.

143. [Moreda et al., 2008b] Moreda, P., Llorens, H., Saquete, E., and Palomar, M. (2008b). The influence of semantic roles in qa: a comparative analysis. In *Proceedings of the SEPLN 2008*, vol.41, pp.55-62, Madrid, Spain.

144. [Moreda, 2008] Moreda, P. (2008). Los Roles Semánticos en la Tecnología del Lenguaje Humano: Anotación y Aplicación. Doctoral Thesis. University of Alicante.

145. [Mullen and Collier, 2004] Mullen, T. and Collier, M. (2004). Sentiment Analysis Using Support Vector Machines with Diverse Information Sources. In *Proceedings of EMNLP 2004.*

146. [Mullen and Malouf, 2006] Mullen, T. and Malouf, R. (2006). A preliminary investigation into sentiment analysis of informal political discourse. In *Proceedings of the AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW),* 2006: 159-162.

147. [Nasukawa and Yi, 2003] Nasukawa, T. and Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the Conference on Knowledge Capture (K-CAP)*, 2003: 70-77.

148. [Neviarouskaya et al., 2010] Neviarouskaya, A., Prendinger, H. and Ishizuka, M. (2010). EmoHeart: Conveying Emotions in Second Life Based on Affect Sensing from Text. *Advances in Human-Computer Interaction, vol. 2010.*

149. [Ng et al., 2006] Ng, V., Dasgupta, S., and Arifin, S.M. N. (2006). Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In *Proceedings 40th*

Annual Meeting of the Association for Computational Linguistics, 2006: 611-618.

150. [Niu et al., 2005] Niu, Y., Zhu, X., Li, J. and Hirst, G. (2005). Analysis of polarity information in medical text. In *Proceedings of the American Medical Informatics Association 2005 Annual Symposium.*

151. [Oatley et al., 2006] Oatley, K., Keltner, D. and Jenkins, J. M. (2006). *Understanding Emotions*. Wiley-Blackwell, 2nd edition.

152. [Oatley, 2004] Oatley, K. (2004). *Emotions. A Brief History*. Oxford: Wiley Blackwell.

153. [Ortony et al., 2005] Ortony, A., Norman, D., and Revelle, W. (2005). Affect and pro-affect in effective functioning. In *J.M. Fellous & M.A. Arbib, Who needs emotions: The brain meets the machine*. New York: Oxford University Press.

154. [Ortony, 1997] Ortony, A. (1997). Metaphor and Thought. Cambridge Press, 2nd edition.

155. [Ounis et al., 2007] Ounis, I., Macdonald, C., de Rijke, M., Mishne, G., and Soboroff, I. (2007). Overview of the TREC 2006 blog track. In *The Fifteenth Text REtrieval Conference (TREC 2006),* National Institute for Science and Technogies Publication, NIST, Gaisthersbury, MD, USA.

156. [Pang and Lee, 2003] Pang, B. and Lee, L. (2003). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the ACL,* 2003: 115–124.

157. [Pang and Lee, 2004] Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL 2004,* pp.271-278, Barcelona, Spain.

158. [Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval,* Vol 2, Nr. 1-2, 2008.

159. [Pang et al., 2002] Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02*, 2002: 79-86.

160. [Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing,* 2002: 79-86, Philadelphia, PA, USA.

161. [Pantel and Ravichandran, 2004] Pantel, P. and Ravichandran, D. (2004). Automatically Labeling Semantic Classes. In *Proceedings of HLT/NAACL-04*, pp. 321-328. Boston.

162. [Parrot, 2001] Parrott, W. (2001). Emotions in Social Psychology. Psychology Press, Philadelphia.

163. [Pennebaker et al., 2003] Pennebaker, J. W., Mehl, M. R. and Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. In *Annual Review of Psychology* 54, 547-577.

164. [Piao et al., 2007] Piao, S., Ananiadon, S., Tsuruoka, Y., Sasaki, Y., and McNaught, J. (2007). Mining opinion polarity of citations. In *Proceedings of the International Workshop on Computational Semantics IWCS 2007*.

165. [Picard, 1995] Picard, R. (1995). Affective computing. Technical report, MIT Media Laboratory.

166. [Picard, 1997] Picard, R. (1997). Affective Computing. MIT Press.

167. [Platt, 1998] Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Microsoft Research Technical Report MSR-TR-98-14.

168. [Plutchik, 2001] Plutchik, R. (2001). The Nature of Emotions. *American Scientist*. 89, 344.

169. [Polanyi and Zaenen, 2004] Polanyi, L. and Zaenen, A. (2004). Exploring attitude and affect in text: Theories and applications. Technical Report SS-04-07.

170. [Popescu and Etzioni, 2005] Popescu, A. M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *In Proceedings of HLT-EMNLP 2005*, Vancouver, Canada.

171. [Pouliquen et al., 2007] Pouliquen, B., Steinberger, R., and Best, C. (2007). Automatic detection of quotations in multilingual news. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2007)*, pp.25-32, Borovets, Bulgaria.

172. [Pustejovsky and Wiebe, 2005] Pustejovsky, J. and Wiebe, J. (2005). Introduction to special issue on advances in question answering. *Language Resources and Evaluation* (2005), (39).

173. [Quirk, 1985] Quirk, R. (1985). *A Comprehensive Grammar of the English Language*. Longman, London, UK.

174. [Ramchurn, 2004] Ramchurn, S. (2004). Multi-Agent Negotiation Using Trust and Persuasion. Ph.D Thesis, University of Southhampton, UK.

175. [Ratner, 2000] Ratner, C. (2000). A cultural-psychological analysis of emotions. *Culture and Psychology*, vol.(6), pp.5-39.

176. [Riloff and Wiebe, 2003] Riloff, E. and Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In Proceedings of the 2003 *Conference on Empirical Methods in Natural Language Processing, EMNLP 2003,* pp.105-112.

177. [Riloff et al., 2003] Riloff, E., Wiebe, J., and Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Conference on Natural Language Learning (CoNLL) 2003,* pp.25-32, Edmonton, Canada.

178. [Riloff et al., 2005] Riloff, E., Wiebe, J., and Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. In *Proceedings of the AAAI 2005.*

179. [Rumbell et al., 2008] Rumbell, T., Barnden, J., Lee, M., and Wallington, A. (2008). Affect in metaphor: Developments with wordnet. In *Proceedings of AISB Symposium of Affect in Human and Machine*.

180. [Saggion and Funk, 2010] Saggion, H., Funk, A. (2010). Interpreting SentiWordNet for opinion classification. In *Proceedings of LREC 2010,* RTF, Tagged XML, Bib Tex Google Scholar.

181. [Saggion et al., 2010] Saggion, H., Lloret, E., Palomar, M. (2010). Using text summaries for predicting rating scales. In: *Proceedings of the 1st Workshop on Subjectivity and Sentiment Analysis (WASSA 2010),*pp.44-51.

182. [Saquete et al., 2006] Saquete, E., Ferrández, O., Martínez-Barco, P. and Muñoz, R. (2006). Reconocimiento temporal para el italiano combinando técnicas de aprendizaje automático y adquisición automática de conocimiento. In Proceedings of the 22nd International Conference of the Spanish Society for the Natural Language Processing (SEPLN 2006), pp. 161-168.

183. [Scherer and Walbott, 1997] Scherer, K. and Wallbott, H. (1997). *The ISEAR Questionnaire and Codebook*. Geneva Emotion Research Group.

184. [Scherer, 1987] Scherer, K. (1987). Toward a dynamic theory of emotion. The component process of affective states. *Cognition and Emotion,* 1(1), 1-98.

185. [Scherer, 1989] Scherer, K. (1989). Appraisal Theory. Handbook of Cognition and Emotion. John Wiley & Sons Ltd.

186. [Scherer, 1993] Scherer, K. R. (1993). Studying the Emotion-Antecedent Appraisal Process: An Expert System Approach. *Cognition and Emotion, 7 (3/*4), 323-355.

187. [Scherer, 2001] Scherer, K. (2001). Emotional expression is subject to social and technological change: Extrapolating to the future. *Social Science Information*, 40(1), 125-151.

188. [Scherer, 2005] Scherer, K. (2005). What are emotions? and how can they be measured? *Social Science Information*, 3(44), 695-729.

189. [Shen et al., 2007] Shen, D., Wiegand, M., Merkel, A., Kazalski, S., Hunsicker, S., Leidner, J. L., and Klakow, D. (2007). The Alyssa System at TREC QA 2007: Do we need blog06? In *Proceedings of TREC 2007*.

190. [Solomon, 2005] Solomon, R. C. (2005). Subjectivity. In Honderich, Ted. Oxford Companion to Philosophy, Oxford University Press, p.900.

191. [Somasundaran et al., 2007] Somasundaran, S., Wilson, T., Wiebe, J., and Stoyanov, V. (2007). Qa with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007),* Boulder< Colorado, USA.

192. [Sperber and Wilson, 1995] Sperber, D. and Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.) Oxford: Blackwell. Also: Sperber, D. and Wilson, D. (2004) "Relevance Theory" in G. Ward and L. Horn (eds) *Handbook of Pragmatics*. Oxford: Blackwell, 607-632.

193. [Steinberger and Ježek, 2004] Steinberger, J. and Ježek, K.(2004). Text summarization and singular value decomposition. In Lecture Notes in Computer Science, Nr. 3261, pp. 245-249, Proceedings of the 3rd Advances in Information Systems (ADVIS) conference, Izmir, Turkey.

194. [Steinberger and Ježek, 2009] Steinberger, J. and Ježek, K.(2009). Update summarization based on novel topic distribution. In Proceedings of the 9th ACM DocEng, Munich, Germany.

195. [Steinberger et al., 2007] Steinberger, J., Poesio, M., Kabadjov, M. A., and Ježek, K. (2007). Two uses of anaphora resolution in *Summmarization. Information Processing and Management,* 43(6):1663–1680. Special Issue on Text Summarisation (Donna Harman, ed.).

196. [Stevenson, 1963] Stevenson, C. (1963). *Facts and Values: Studies in Ethical Analysis*. Yale University Press, New Haven, USA.

197. [Stone et al., 1966] Stone, P., Dumphy, D. C., Smith, M. S., and Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge, USA.

198. [Stoyanov and Cardie, 2006] Stoyanov, V. and Cardie, C. (2006). Toward opinion summarization: Linking the sources. *In Proceedingns of the COLING-ACL 2006 Workshop on Sentiment and Subjectivity in Text.*

199. [Stoyanov et al., 2004] Stoyanov, V., Cardie, C., Lawrence, D. and Wiebe, J. (2004). Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications,* AAIT Press.

200. [Stoyanov et al., 2005] Stoyanov, V., Cardie, C., and Wiebe, J. (2005). Multiperspective question answering using the opqa corpus. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005),* pp.923-930, Vancouver, Canada.

201. [Strapparava and Mihalcea, 2007] Strapparava, C. and Mihalcea, R. (2007). Semeval 2007 task 14: Affective text. In *Proceedings of ACL 2007,* Prague, Czech Republic.

202. [Strapparava and Valitutti, 2004] Strapparava, C. and Valitutti, A. (2004). Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004),* pp.1083-1086, Lisbon, Portugal.

203. [Studer et al., 1998] Studer, R., Benjamins, R. V. and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data and Knowledge Engineering*, 25(1-2):161–197.

204. [Subasic and Huettner, 2000] Subasic, P. and Huettner, A. (2000). Affect Analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy System, 9, 483-496.*

205. [Suchanek et al., 2007] Suchanek, F., Kasnei, G. and Weikum, G. (2007). YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *Proceedings of WWW 2007,* pp.697-706, Banff, Alberto, USA.

206. [Taboada and Grieve, 2004] Taboada, M. and Grieve, J. (2004) Analyzing Appraisal Automatically. *American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text*. Stanford. March 2004. AAAI Technical Report SS-04-07, pp.158-161.

207. [Tan and Zhang, 2007] Tan, S. and Zhang, J. (2007). An empirical study of sentiment analysis for chinese documents. In *Expert Systems with Applications.*

208. [Tanev et al., 2009] Tanev, H., Pouliquen, B., Zavarella, B., and Steinberger, R. (2009). Automatic expansion of a social network using sentiment analysis, *Annals of Information Systems*, 2009, Springer, Verlag.

209. [Tanev, 2007] Tanev, H. (2007). Unsupervised learning of social networks from a multiple-source news corpus. In *Proceedings of the Workshop Multi-source Multilingual Information Extraction and Summarization (MMIES 2007)*, held at RANLP 2007, Borovets, Bulgaria.

210. [Terveen et al., 1997] Terveen, L., Hill, W., Amento, B., McDonald, D. and Creter, J. (1997). PHOAKS: A system for sharing recommendations. *Communications of the Association for Computing Machinery (CACM)*, 40(3):59–62, 1997.

211. [Teufel and Moens, 2000] Teufel, S. and Moens, M. (2000). What's yours and what's mine: Determining intellectual attribution in scientific texts. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

212. [Thayer, 2001] Thayer, R. E. (2001). Calm Energy: How People Regulate Mood With Food and Exercise. Oxford University Press, N.Y.

213. [Thomas et al., 2006] Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of EMNLP 2006*, pp,327-335.

214. [Tong, 2001] Tong, R.M. (2001). An operational system for detecting and tracking opinions in on-line discussion. In *Proceedings of the Workshop on Operational Text Classification (OTC)*, 2001, New Orleans, USA.

215. [Toprak et al., 2010] Toprak, C., Jakob, N. and Gurevych, I. (2010). Sentence and Expression Level Annotation of Opinions in User-Generated Discourse. In *Proceedings of ACL* 2010: 575-584.

216. [Turing, 1950] Turing, A. (1950). Computing machinery and intelligence. Mind, 1950.

217. [Turney and Littman, 2003] Turney, P. and Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4), 315-346.

218. [Turney, 2002] Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings 40th Annual Meeting of the Association for Computational Linguistics, ACL 2002*, 417-424, Philadelphia, USA.

219. [Uspensky, 1973] Uspensky, B. (1973). A Poetics of Composition. University of California Press, Berkeley, California.

220. [Van den Bos, 2006] Van den Bos, G. (2006). APA Dictionary of Psychology. Washington, DC: American Psychological Association.

221. [Varma et al., 2008] Varma, V., Pingali, P., Katragadda, R., Krisha, S., Ganesh, S., Sarvabhotla, K., Garapati, H., Gopisetty, H., Reddy, V., Bysani, P., and Bharadwaj, R. (2008). Iit hyderabad at tac 2008. *In Proceedings of the Text Analysis Conference (TAC) 2008.*

222. [Vázquez et al., 2004] Vázquez, S., Montoyo, A., and Rigau, G. (2004). Using relevant domains resource for word sense disambiguation. In *Proceedings of the International Conference on Artificial Intelligence*, IC-AI, 2004: 784-789.

223. [Whitelaw et al., 2005] Whitelaw, C., Garg, N., and Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the CIKM* 2005: 625-631, Bremen, Germany.

224. [Wiberg, 2004] Wiberg, M.(ed) (2004). *The Interaction Society: Theories, Practice and Supportive Technologies*. Information Science Publishing, Idea Group Inc.

225. [Wiebe and Mihalcea, 2006] Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity. In *Proceedings of ACL 2006.*

226. [Wiebe and Riloff, 2005] Wiebe, J. and Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-05)*, 73-99.

227. [Wiebe and Wilson, 2005] Wiebe, J. and Wilson, T. (2005). Annotating attribution and private states. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky.*

228. [Wiebe et al., 2001] Wiebe, J., Bruce, R. , Bell, M., Martin, M. and Wilson, T. (2001). A Corpus Study of Evaluative and Speculative Language. In *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-2001).* Aalborg, Denmark, pp.186–195.

229. [Wiebe et al., 2004] Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004). Learning subjective language, *Computational Linguistics*, 30(3), 277-308.

230. [Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, vol. 39(2-3),119-122.

231. [Wiebe, 1994] Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2), 233-287..

232. [Wiegand et al., 2010] Wiegand, M., Balahur, A., Roth, B., Klakow, D., Montoyo, A. (2010). A Survey on the Role of Negation in Sentiment Analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP 2010.*

233. [Wilson and Wiebe, 2003] Wilson, T. and Wiebe, J. (2003). Annotating opinions in the world press. In *Proceedings of SIGdial 2003.*

234. [Wilson et al., 2004] Wilson, T., Wiebe, J., and Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of AAAI 2004*,761-769, San Jose, USA.

235. [Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP 2005*, pp.347-354, Vancouver, Canada.

236. [Wu and Jin, 2010] Wu, Y., Jin, P. (2010). SemEval-2010 Task 18: Disambiguating Sentiment Ambiguous Adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 81–85, Uppsala, Sweden, 15-16 July 2010.

237. [Yi et al., 2003] Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the IEEE International Conference on Data Mining (ICDM).* 2003: 427-434.

238. [Yu and Hatzivassiloglou, 2003] Yu, D. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP),* 2003: 129-136, Shapporo, Japan.

ANNEX A. SCIENTIFIC CONTRIBUTIONS AND MERITS DURING THE PHD STUDIES

1. SCIENTIFIC PUBLICATIONS:

BOOK CHAPTERS:

1. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *OpAL: a System for Mining Opinion from Text for Business Applications.* In Marta E. Zorrilla, Jose-Norberto Mazón, Óscar Ferrández, Irene Garrigós, Florian Daniel and Juan Trujillo (Eds.). Business Intelligence Applications and the Web: Models, Systems and Technologies, IGI Global Press, 2010 (in press).
2. Balahur A.; Hermida, J. *Lost in Translation: The Impact of New Technologies on the Education Gap.* In Balahur, D., Qvarsel, B. (Eds.) Children's Rights to Education and Information in a Globalized World. Alexandru Ioan Cuza University Press, pp. 103-116, ISBN 978-973-703-374-1, 2008.

JOURNAL ARTICLES:

1. Balahur, A.; Hermida J.M.; Montoyo, A. *Detecting Implicit Expressions of Emotion Using Commonsense Knowledge and Appraisal Models.* IEEE Transactions on Affective Computing. Special Issue on Naturalistic Affect Resources for System Building and Evaluation, 2011 (accepted).
2. Balahur, A.; Hermida J.M.; Montoyo, A. *Building and Exploiting EmotiNet: a Knowledge Base for Emotion Detection Based on the Appraisal Theory Model.* In Lecture Notes in Computer Science Nr. 6677, 2011 - Proceedings of the 9th International Symposium on Neural Networks (in press).
3. Balahur, A.; Hermida J.M.; Montoyo, A. *EmotiNet: a Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories*. In Lecture Notes in Computer Science – Proceedings of the 16th International Conference on the Applications of Natural Language to Information Systems, 2011 (in press).
4. Balahur, A.; Kabadjov, M.; Steinberger, J.; Steinberger, R.; Montoyo, A. *Opinion Summarization for Mass Opinion Estimation.* Journal of Intelligent Information Systems, 2011 (accepted).
5. Kabadjov, M.; Balahur, A.; Boldrini, E. Sentimtent Intensity: Is It a Good Summary Indicator? In Lecture Notes in Artificial Intelligence, Nr. 6562 –

Selected Papers from the 4th Language Technology Conference 2009 (in press).

6. Balahur, A., Kabadjov, M. Steinberger, J. *Exploiting Higher-level Semantic Information for the Opinion-oriented Summarization of Blogs.* International Journal of Computational Linguistics and Applications, ISSN: 0976-0962, Vol. 1, No. 1-2, pp. 45-59, 2010.
7. Martín, T.; Balahur, A.; Montoyo, A.; Pons, A. *Word Sense Disambiguation in Opinion Mining: Pros and Cons.* In Journal Research in Computing Science, Special Issue: Natural Language Processing and its Applications, ISSN: 1870-4069, Vol. 46, pp. 119-130.
8. Balahur, A.; Kozareva, Z.; Montoyo, A. *Determining the Polarity and Source of Opinions Expressed in Political Debates.* In Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing) 2009 (Lecture Notes in Computer Science No. 5449, 2009).
9. Balahur, A.; Montoyo, A. *A Semantic Relatedness Approach to Classifying Opinion from Web Reviews*. In Procesamiento del Lenguaje Natural No. 42, 2009.
10. Balahur, A; Balahur, P. *What Does the World Think About You? Opinion Mining and Sentiment Analysis in the Social Web*. In "Analele Stiintifice ale Universitatii Al.I. Cuza Iasi", Nr.2, pp. 101-110, 2009, ISSN: 2065-2917.
11. Balahur, A.; Montoyo, A. *Definición de disparador de emoción asociado a la cultura y aplicación a la clasificación de la valencia y de la emoción en textos.* Procesamiento del Lenguaje Natural, Vol: 40, Num: 40, 2008.
12. Balahur, A.; Montoyo, *A. Multilingual Feature-Driven Opinion Extraction and Summarization from Customer Reviews.* Journal Lecture Notes in Computer Science, 5029, pp. 345-346 – Proceedings of the 13th International Conference on Applications of Natural Language to Information Systems (NLDB 2008).

ARTICLES IN CONFERENCE PROCEEDINGS:

1. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *Going Beyond Traditional QA Systems: Challenges and Keys in Opinion Question Answering.* In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010).
2. Wiegand, M.; Balahur, A., Roth, B.; Klakow, D., Montoyo, A. *A Survey on the Role of Negation in Sentiment Analysis.* In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP 2010.

3. Balahur, A.; Montoyo, A. *OpAL: Applying Opinion Mining Techniques for the Disambiguation of Sentiment Ambiguous Adjectives in SemEval-2 Task 18.* In Proceedings of SemEval-2, the 5th International Workshop on Semantic Evaluation, satellite workshop to ACL 2010.
4. Boldrini, E.; Balahur, A., Martínez-Barco, P.; Montoyo, A. *EmotiBlog: a finer-grained and more precise learning of subjectivity expression models*. In Proceedings of the 4th Linguistic Annotation Workshop (LAW IV), satellite workshop to ACL 2010.
5. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *The OpAL System at NTCIR 8 MOAT*. In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR 8), 2010.
6. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *Opinion Question Answering: Towards a Unified Approach*. In Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010).
7. Balahur, A.; Steinberger, R.; Kabadjov, M.; Zavarella, V.; Van der Goot, E.; Halkia, M.; Pouliquen, B.; Belyaeva, J. *Sentiment Analysis in the News*. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 May 2010.
8. Martín-Wanton, T.; Pons-Porrata, A.; Montoyo-Guijarro, A.; Balahur, A. *Opinion Polarity Detection – Using Word Sense Disambiguation to Determine the Polarity of Opinions*. In Proceedings of the 2ndInternational Conference on Agents and Artificial Intelligence (ICAART 2010).
9. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *Opinion and Generic Question Answering systems: a Performance Analysis. In Proceedings of the ACL-IJCNLP 2009 Conference.*
10. Balahur, A.; Kabadjov, M.;Steinberger, J.; Steinberger, R.; Montoyo, A. *Summarizing Opinions in Blog Threads*. Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC), pp. 606-613, Hong Kong, 3-5 December 2009.
11. Kabadjov, M.; Balahur, A.; Boldrini, E. *Sentiment Intensity: Is It a Good Summary Indicator?*.Proceedings of the 4th Language Technology Conference LTC, pp. 380-384. Poznan, Poland, 6-8.11.2009.
12. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *Evaluating Question Answering Systems on Opinión Queries. Proceedings of TSA 2009 (The First CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement).*

13. Balahur, A.; Steinberger, R. *Rethinking Opinion Mining in Newspaper Articles: from Theory to Practice and Back. Proceedings of the first workshop on Opinion Mining and Sentiment Analysis (WOMSA 2009).*
14. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *A comparative study of Open Domain and Opinion Question Answering Systems for Factual and Opinionated Queries.RANLP 2009.*
15. Balahur, A.; Lloret, E.; Boldrini, E.; Montoyo, A.; Palomar, M; Martínez-Barco, P. *Summarizing Threads in Blogs Using Opinion Polarity. Proceedings of the EETTs 2009 Workshop, satellite to the RANLP 2009 Conference.*
16. Balahur, A.; Steinberger, R.; van der Goot, E.; Pouliquen, B.; Kabadjov, M. *Opinion Mining from Newspaper Quotations. In Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content, 2009 IEEE/WIC/ACM International Conference on Web Intelligence held in conjunction with IAT'09, September 2009, Milan, Italy.*
17. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *A Comparative Study of Open Domain and Opinion Question Answering Systems for Factual and Opinionated Queries. In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP 2009).*
18. Balahur, A.; Montoyo, A. *Semantic Approaches to Fine and Coarse-Grained Feature-Based Opinion Mining. In Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems (NLDB 2009).*
19. Boldrini, E.; Balahur, A., Martínez-Barco, P.; Montoyo, A. *EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. In Proceedings of the 5th International Conference on Data Mining (DMIN 2009).*
20. Lloret, E.; Balahur, A.; Palomar, M.; Montoyo, A. *Towards Building a Competitive Opinion Summarization System: Challenges and Keys. In Proceedings of the NAACL-HLT 2009 Student Research Workshop, 2009.*
21. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *Fact Versus Opinion Questions Classification and Answering: Challenges and Keys. In Proceedings of the International Conference on Artificial Intelligence (ICAI 2009).*
22. Balahur, A.; Boldrini, E.; Montoyo, A.; Martínez-Barco, P. *Cross Topic Opinion Mining for Real-time Human-Computer Interaction (Application of EmotiBlog – an Annotation Schema for Emotion Detection). In Proceedings of the Workshop on Natural Language Processing and Cognitive Science (NLPCS 2009).*

23. Balahur, A.; Lloret, E.; Ferrández, O.;Montoyo, A., Palomar, M.; Muñoz, R. *The DLSIUAES Team's Participation in the TAC 2008 Tracks Proceedings of the Text Analysis Conference 2008 Workshop, Washington, USA, 2008.*
24. Iftene, A.; Balahur, A. *A Language Independent Approach for Recognizing Textual Entailment. Research in Computing Science , Vol: 334, 2008.*
25. Balahur, A.; Montoyo, A. *Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification. In Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine, Aberdeen, Scotland, 2008.*
26. Balahur, A.; Montoyo, A. *Building a Recommender System Using Community Level Social Filtering. In Proceedings of the 5th International Workshop on Natural Language and Cognitive Science, Barcelona, Spain, 2008.*

2. PARTICIPATION IN SCIENTIFIC COMPETITIONS:

1. Participation in the competition SemEval 2010, task 18 – "Disambiguation of sentiment ambiguous adjectives" (in Traditional Chinese). (Rank: 5/16)
2. Participation in the competition NTCIR 8 MOAT (Multilingual Opinion Analysis Task), English mono-lingual task and English-Chinese cross-lingual task. (Rank: 1/6 cross-lingual)
3. Participation in the competition Opinion Pilot in the Text Analysis Conference (TAC) 2008. (5/18 in F-measure)
4. Participation in the competition Answer Validation Exercise (AVE) 2008 in Romanian and English, at CLEF 2008 (Cross-Language Evaluation Forum). (Rank: 1/8 Romanian)
5. Participation in the competition Answer Validation Exercise (AVE) 2007 in English, at CLEF 2007. (Rank: 1/13)
6. Participation in the competition Recognizing Textual Entailment 3 (RTE 3), organized by the PASCAL Excellence Network. (Rank: 3/26)

3. PARTICIPATION IN PROJECTS:

1. **Title of the project:** Living in Surveillance Societies (LiSS) COST (European Cooperation in the field of Scientific and Technical Research) Action IS0807 (European Commission – FP7)

2. **Title of the project:** Las tecnologías del lenguaje humano ante los nuevos retos de la comunicación digital ACOMP/2010/286 – Consellería de Educación
3. **Title of the project:** Procesamiento del lenguaje y sistemas de información (GPLSI) VIGROB-100 – Universidad de Alicante
4. **Title of the project:** Las Tecnologías del Lenguaje Humano ante los nuevos retos de la comunicación digital (TIN2009-13391-C04-01) - Ministerio de Ciencia e Innovación (Ministry of Science and Innovation)
5. **Title of the project:** Desarollo conjunto de un grupo de herramientas y recursos de tecnologías del lenguaje humano (Joint Development of Tools and Resources for Human Language Technologies)
6. **Title of the project:** Desarollo y licencia de software para la extracción de información de documentos notariales (Software Development and Licencing for the Extraction of Information from Notary Documents)
7. **P ometeo /2009/119 -** Desarrollo de técnicas inteligentes ería de textos.

4. PARTICIPATION IN PROGRAM COMMITTEES:

I was a Program Committee member of the following events:

- eETTs 2009 (Workshop on Events in Emerging Text Types), with RANLP 2009;
- WOMSA 2009 (Workshop on Opinion Mining and Sentiment Analysis), with CAEPIA 2009;
- SMUC 2010 (2nd International Workshop on Search and Mining User-generated Contents), with CIKM 2010;
- WASSA 2010 (1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis), with ECAI 2010;
- WASSA 2.011 (2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis), with ACL-HLT 2011;
- SNMART 2011 (Social Network Mining, Analysis and Research Trends: Techniques and Applications);
- SAAIP 2011 (Sentiment Analysis where AI Meets Psychology), with IJCNLP 2011;
- SENTIRE 2011 (Sentiment Elicitation from Natural Text for Information Retrieval and Extraction), with ICMD 2011.
- ICL 2011: Workshop on Iberian Cross-Language NLP tasks at SEPLN 2011.
- RANLP 2011 – Conference on Recent Advances in Natural Language Processing.

I was also a reviewer for the Language Resources and Evaluation journal - Special Issue on Short Texts, IEEE Transactions on Affective Computing - Special Issue on Naturalistic Affect Resources for System Building and Evaluation, ACM Transactions on Itelligent Systems and Technology, RANLP 2009, NLDB 2009, SEPLN 2009, IceTAL 2010, NLDB 2011.

5. ORGANIZATION OF SCIENTIFIC EVENTS:

Together with Prof. Andrés Montoyo, Prof. Patricio Martínez-Barco and Ester Boldrini, I was an organizer (co-chair) of the WASSA 2010 (1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis), celebrated with the 19th European Conference on Artificial Intelligence (ECAI 2010). This year, we are organizing the second edition of the workshop - WASSA 2.011 at ACL-HLT 2011, in Portland, Oregon, on the 24th of June. Additionally, I am one of the organizers of NLDB 2011 (16th International Conference on the Applications of Natural Language to Information Systems).

6. AWARDS AND SCHOLARSHIPS OBTAINED:

1. Scholarship from the Ministry of Education for a 3-month research traineeship at a foreign centre, with the aim of obtaining the European Mention on the PhD title, 2009-2010, at the University of Saarland, Saarbrücken, Germany.

2. PhD Scholarship from the University of Alicante (2007-2011).

3. Scholarship from CLEF for the participation in the summer school entitled "TrebleCLEF Summer School on Multilingual Information Access", celebrated between the 15th and 19th of June 2009 in Santa Croce de Fossabanda, Pisa, Italy. (summer school director - Prof. Carol Peters) .

4. 6-month traineeship with the European Commission, Joint Research Centre, Ispra, Italy. Institute for the Protection and Security of the Citizens, Global Security and Crisis Management Unit, Open Source Text Information Mining and Analysis Action.

5. Scholarship from the Duques de Soria Foundation (Fundación Duques de Soria) for the participation in the language technologies summer courses, entitled "Técnicas de extracción y visualización de información: aplicación en la construcción de portales especializados", 7-11 July 2008, Soria, Spain.

ANNEX B. RESUMEN EN CASTELLANO

INTRODUCCIÓN Y MOTIVACIÓN

La era en la que vivimos ha sido denominada de muchas maneras. "Aldea global", "era tecnotrónica", "sociedad postindustrial", "sociedad de la información", "era de la información", y "sociedad del conocimiento" son sólo algunos de los términos utilizados en un intento de describir los profundos cambios que han ocurrido en la vida de las personas y las sociedades en todo el mundo como resultado del rápido desarrollo de las tecnologías de la información y la comunicación (TIC), el acceso a Internet y su transformación gradual en una Web social. En este nuevo contexto, tener acceso a grandes cantidades de información ya no es un problema. Todos los días cualquier persona con una conexión a Internet tiene acceso a terabytes de nueva información que se producen en la Web. A diferencia de otros tiempos, cuando la búsqueda de fuentes de información era el problema principal de los usuarios, la sociedad de la información reta a las compañías y los individuos a crear y utilizar mecanismos para la búsqueda, recuperación y procesamiento de datos relevantes entre la gran cantidad de información disponible, para obtener conocimiento que puedan utilizar en su beneficio. En contraste con unos años atrás, cuando esta ventaja dependía de la capacidad de encontrar fuentes de información, en la sociedad actual, que está inundada por un flujo de información que cambia a un ritmo alto, la ventaja viene dada por la calidad (exactitud, fiabilidad) del conocimiento extraído y su concreción. En la época en que vivimos, la información se ha convertido en el principal objeto de comercio. Tener a mano información oportuna de alta calidad es crucial en todos los ámbitos de actividad humana: social, político y económico, por mencionar sólo algunos.

Sin embargo, en muchos casos, la información pertinente no se encuentra en fuentes estructuradas (es decir, tablas o bases de datos), sino en documentos no estructurados, escritos en lenguaje humano. La elevada cantidad en la que estos datos se encuentran en la web requiere el uso de técnicas de procesamiento automático para poder tratarlos.

La disciplina que se ocupa del tratamiento automático del lenguaje natural (o lenguaje humano) en textos o el habla se llama Procesamiento del Lenguaje Natural (PLN). El PLN es parte del área de investigación de la Inteligencia Artificial (IA), que se define como "la ciencia y la ingeniería de hacer máquinas inteligentes" (McCarthy, 1959) mediante la simulación de los mecanismos de la inteligencia humana. Dentro del PLN existen diversas áreas de investigación. Cada una de estas áreas se ha establecido para resolver problemas difíciles que se pueden encontrar en cualquier ámbito del PLN (por ejemplo, la Desambiguación del Sentido de las Palabras, la Resolución de la Co-referencia, la Resolución de las Expresiones Temporales, el Etiquetado de Roles Semánticos), o en dependencia de una aplicación final específica (por ejemplo, la Recuperación de Información, la Extracción de Información, la Búsqueda de Respuestas, los Resúmenes de Texto, la Traducción Automática).

Tradicionalmente, estas áreas de aplicación de la PLN fueron diseñadas para el tratamiento de los textos que describen datos factuales (hechos que se pueden observar y comprobar en la realidad). No obstante, hoy en día, la información sobre hechos ya no es la fuente principal de donde se extrae el conocimiento fundamental o más básico.

El presente está marcado por la creciente influencia de la web social (la web de la interacción y la comunicación) en las vidas de las personas en todo el mundo. Ahora, más que nunca, la gente está totalmente dispuesta y feliz por compartir sus vidas, conocimientos, experiencias y pensamientos con el mundo entero, a través de blogs, microblogs, foros, wikis o, incluso, sitios de comercio electrónico que dan la opción de compartir reseñas sobre los productos que venden. La gente está participando activamente en los principales acontecimientos que tienen lugar en todas las esferas de la sociedad, expresando su opinión sobre ellos y comentando las noticias que aparecen. El gran volumen de datos que contienen opiniones disponibles en Internet, en reseñas, foros, blogs, microblogs y redes sociales ha producido un importante cambio en la forma en que las personas se comunican, comparten conocimientos y emociones e influyen en el comportamiento social, político y económico. En consecuencia, esta nueva realidad ha dado lugar a importantes transformaciones en la forma, extensión y rapidez de circulación de las noticias y sus opiniones asociadas, dando lugar a fenómenos sociales, económicos y psicológicos nuevos y desafiantes.

Para estudiar estos fenómenos y abordar la cuestión de extraer el conocimiento fundamental que en la actualidad figura en los textos que contienen expresiones con sentimientos, han nacido nuevos campos de investigación dentro del PLN, que tienen el objetivo de detectar la subjetividad en el texto y/o extraer y clasificar los sentimientos de las opiniones expresadas en distintas categorías (por lo general en positivos, negativos y neutrales). Las nuevas tareas que se abordan en el PLN son principalmente el ***análisis de la subjetividad*** (que trata sobre los "estados privados" (Banfield, 1982), un término que contiene sentimientos, opiniones, emociones, evaluaciones, creencias y especulaciones), ***el análisis de sentimientos*** y la ***minería de opiniones.*** Esta no ha sido la única forma de referirse a los enfoques adoptados. También se han utilizado también otras terminologías que utilizaban otros términos para denominar a las tareas, por ejemplo, minería de reseñas o la extracción de valoración. Asimismo, los términos de “análisis de sentimientos” y “minería de opiniones” se han utilizado indistintamente, ya que algunos autores consideran que hacen referencia a la misma tarea (Pang and Lee, 2008). Uno de los problemas del PLN estrechamente relacionado también con las tareas mencionadas es la “detección de la emoción”, que tiene como objetivo la clasificación de los textos de acuerdo con la emoción expresada en ellos. Todas estas áreas de investigación son parte de una esfera más amplia de la Inteligencia Artificial denominada ***computación del afecto*** (Picard, 1995).

La presente tesis doctoral se ocupa de las cuestiones y los desafíos en el desarrollo de métodos y recursos para la tarea del PLN denominada ***análisis de sentimientos***. Definido de forma general, el objetivo de esta tarea es la detección automática de los sentimientos expresados en textos (normalmente por una fuente, sobre un “objeto”, que puede ser una persona, un evento, un producto, una organización etc.) y su clasificación según la polaridad/orientación que tienen (normalmente *positiva*, *negativa* o *neutra*, aunque distintos autores han propuesto escalas más finas de sentimientos, incluyendo por ejemplo las clases *muy positivo* o *muy negativo*).

PROBLEMAS DEL DOMINIO Y OBJETIVOS DE LA INVESTIGACIÓN

La investigación en este campo, la minería de opiniones (análisis de sentimientos), ha demostrado que el análisis de los sentimientos es un problema difícil, que se tiene que abordar desde diferentes perspectivas y en diferentes niveles, dependiendo de una serie de factores. Estos factores incluyen: nivel de de interés (general o específico, dependiendo de si la opinión general sobre el objeto en cuestión es suficiente o se necesita conocimiento detallado de los sentimientos expresados sobre distintos componentes del objeto), la fórmula de consulta (“Nokia E65” / “¿Por qué la gente compra el Nokia E65?”), el tipo de texto (revisión en un foro/blog/diálogo/artículo de periódico), y la forma de expresar la opinión - directamente (mediante declaraciones opinión, por ejemplo, “¡Me parece que este producto es maravilloso!” o “¡Esta es una iniciativa brillante!”), de forma indirecta (utilizando vocabulario relacionado con la expresión del afecto, por ejemplo, “¡Me encantan las fotos tomadas con esta cámara!” o “Personalmente, ¡estoy conmocionado por cómo se puede proponer una ley así!”) o implícitamente (con adjetivos que expresan una evaluación, cuyo objeto se sobrentiende – por ejemplo, "Es ligero como una pluma y cabe perfectamente en mi bolsillo" o presentando una situación factual de la que se puede inferir, utilizando conocimiento común, una emoción positiva o negativa – por ejemplo “Se rompió en dos días.”). Otros factores que hacen la tarea de minería de sentimientos difícil es la aplicación final y el tipo de texto que se utiliza (reseñas, que contienen solo opiniones sobre un producto, escritas por una solo fuente, en comparación con blogs o debates, que tienen una estructura de dialogo, en el que se expresan opiniones sobre distintos objetos, por distintas fuentes). Finalmente, para las aplicaciones finales, el análisis de sentimientos no es la primera ni la última tarea que se debe realizar. Para extraer el sentimiento de textos, primero es necesario recuperar un conjunto de documentos relevantes. El resultado del procesamiento de un texto con un sistema de análisis de sentimientos puede tener mucha información redundante e incluso puede no resolver totalmente el problema, debido a la gran cantidad de datos existentes.

Los sistemas implementados para la tarea de análisis de sentimientos se basan en reglas, bolsas de palabras, utilizando un léxico de palabras que tienen una orientación del sentimiento (positivo o negativo), métodos estadísticos o aprendizaje automático.

Analizando los sistemas existentes, hemos identificado los siguientes problemas:

- La tarea de análisis de sentimientos y los conceptos relacionados no son definidos de forma única en los diferentes trabajos de investigación. Por tanto, no está claro siempre si los distintos investigadores que trabajan en el análisis de sentimientos pueden comparar el rendimiento de sus sistemas, ya que los textos sobre los que evalúan pueden tener diferentes elementos anotados.
- La tarea de análisis de sentimientos se resuelve de la misma manera, independientemente del tipo de texto que se procesa y del objetivo de la aplicación final.
- No existen recursos anotados para la tarea de análisis de sentimientos en todos los géneros textuales.
- No existen léxicos de palabras que expresen sentimientos para otros idiomas distintos al inglés.
- La mayoría de sistemas trabajan a nivel léxico, utilizando reglas, léxicos, métodos estadísticos o aprendizaje automático. La investigación que se ha hecho hasta ahora no toma en cuenta otros niveles de análisis, como el sintáctico o semántico. Por tanto, el asegurar que la fuente de la opinión expresada es la requerida o sobre qué objeto se expresa la opinión en un texto son aspectos que no se toman en consideración. Estos aspectos pueden tener un alto impacto sobre el rendimiento y la utilidad de los sistemas de análisis de opiniones.
- La mayor parte de la investigación no distingue sobre los distintos componentes de un texto, en especial sobre el autor, el texto y el lector. La tarea de análisis de sentimientos puede tener diferentes objetivos, dependiendo de la perspectiva que se requiere analizar (por ejemplo, si el autor tiene preferencia sobre un cierto objeto descrito, si el texto contiene

información que es buena o mala en sí, si el lector confía en la fuente de la información).

- Las tareas tradicionales (búsqueda de información, búsqueda de respuestas, resúmenes automáticos) se enfrentan con problemas adicionales en el caso de que la información buscada o resumida es de tipo opinión, dadas por las características del lenguaje afectivo. Por tanto, para poder adecuar los sistemas de este tipo para tratar información que contiene expresiones de afecto, las peculiaridades de este lenguaje tienen que ser estudiadas y se tienen que proponer métodos adecuados para resolver los problemas encontrados de forma eficaz.

El objetivo de nuestro trabajo ha sido crear, explotar y evaluar métodos y recursos tanto nuevos como consagrados para la detección y posterior clasificación de acuerdo a su polaridad (positiva / negativa/ neutro) de los sentimientos expresados en textos.

En concreto, el primer objetivo es desarrollar técnicas adecuadas para la detección y clasificación automática de los sentimientos expresados de forma directa, indirecta o implícita en los textos de diferentes tipos (reseñas, artículos de periódicos, diálogos/debates y blogs) en diferentes idiomas. El segundo objetivo es aplicar los métodos de análisis de sentimientos que se proponen en el contexto o conjuntamente con otras tareas de PLN (búsqueda de respuestas y resúmenes automáticos) y proponer técnicas adecuadas para hacer frente a las cuestiones planteadas en estas tareas por las peculiaridades de la expresión del afecto.

En concreto, nos centramos en:

- Definir la tarea y conceptos generales relacionados, a partir del estudio de las definiciones existentes en la literatura y la clarificación de las inconsistencias detectadas;
- Proponer y evaluar métodos para definir y abordar el análisis de los sentimientos de diversos géneros textuales en diferentes idiomas;
- Redefinir la tarea y proponer métodos para anotar corpus específicos para el análisis de sentimientos en para un tipo de texto en diferentes idiomas, en el caso de que la tarea de análisis de sentimientos no hubiera sido claramente definida para el género textual en cuestión y/o ningún corpus estuviera disponible para el mismo. Estos recursos están disponibles al público para el uso de la comunidad científica;
- Aplicación de técnicas de minería de opinión en el contexto de los sistemas "end-to-end" y también en conjunto con otras tareas del PLN. Para ello, nos hemos concentrado en realizar análisis de emociones en las tareas de búsqueda de respuesta y resumen automático;

- Llevar a cabo experimentos con sistemas de búsqueda de respuesta y sistemas de resúmenes automáticos, diseñados para hacer frente a datos factuales solamente;
- Proponer y evaluar un nuevo marco para lo que llamamos "búsqueda de respuestas a preguntas de opinión" (en inglés - Opinion Question Answering) y los nuevos métodos para "elaborar resúmenes de opiniones de forma automática" (en inglés - Opinion Summarization), tras realizar un conjunto de experimentos que mostraron que los sistemas de búsqueda de respuestas y de resumen automático sobre textos presentando hechos no funcionaban correctamente para analizar textos que contenían opiniones;
- Presentación de un método general para la detección de la emoción expresada de manera implícita en texto. En primer lugar, presentamos el método para construir un léxico de términos que en sí mismos no contienen la emoción, pero que disparan la emoción en un lector. Posteriormente, se propuso un método para resumir textos a partir del análisis de los sentimientos expresados basado en claves lingüísticas, así como se propuso y evaluó un método para representar el texto como las cadenas de acción. La emoción provocada por la situación que se presenta en el texto se juzga posteriormente en base a conocimiento de sentido común sobre el efecto emocional de cada acción en la cadena;
- La evaluación de nuestros enfoques en las competiciones internacionales, a fin de comparar nuestros enfoques con los demás y validarlos.

Con el fin de alcanzar los objetivos propuestos, el trabajo que se presenta ha sido estructurado en torno a responder a cinco preguntas de investigación. Cada uno de los capítulos de esta tesis presenta métodos y evaluaciones hechas con el fin de responder a estas preguntas.

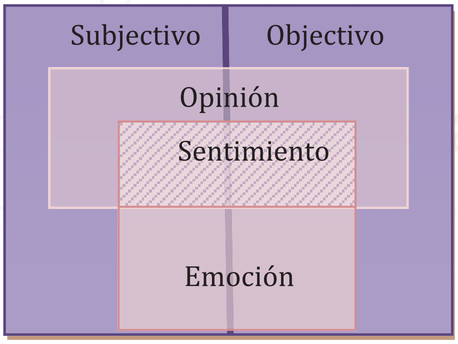
CONTRIBUCIÓN

Este apartado presenta en detalle las contribuciones realizadas a la investigación en el campo de análisis de sentimientos a lo largo de esta tesis y muestra cómo los métodos y recursos propuestos llenan vacíos importantes en la investigación existente. Las principales contribuciones responder a cinco preguntas de investigación:

1. *¿Cómo se puede definir la tarea de análisis de sentimientos y, en una perspectiva más amplia, la minería de opiniones de una manera correcta? ¿Cuáles son los conceptos principales que se deberían definir antes de afrontar la tarea?*

En el capítulo 2, presentamos un conjunto de definiciones asociadas a cada uno de los conceptos involucrados en la tarea de análisis de sentimientos – la subjetividad, la objetividad, la opinión, el sentimiento, la emoción la actitud, y la evaluación.

Nuestra aportación en este capítulo reside en mostrar con claridad que el análisis de los sentimientos y la minería opinión no son sinónimos, aunque en la literatura por lo general son empleados indistintamente. Además, demostramos que la "opinión", como es definida por el diccionario Webster, no es sinónimo de sentimiento. Considerando que los sentimientos son tipos de opiniones, los que reflejan las vivencias (es decir, la parte consciente de las emociones), pero todas las opiniones no son los sentimientos (es decir, hay tipos de opiniones que no son el reflejo de las emociones). También demostramos que el análisis de la subjetividad no está directamente relacionado con el análisis de los sentimientos, a pesar que así está considerado por muchos de los investigadores en el campo. En otras palabras, la detección de frases subjetivas no implica directamente la obtención de las frases que contienen sentimiento. Este último, como expresiones de las evaluaciones basadas en la emoción, no son necesariamente indica en frases subjetivas, pero también puede expresarse en frases objetivas. Las frases subjetivas pueden o no contener expresiones de las emociones. Las relaciones entre los diferentes conceptos se resumen en la siguiente figura, que contiene un diagrama de conjuntos:



Por último, demostramos la existencia de una clara conexión entre el trabajo realizado en el marco de análisis de sentimientos/minería opinión y la de evaluación/análisis de la actitud. A pesar de todas estas áreas se consideran generalmente para referirse al mismo tipo de trabajo, el objetivo más amplio de la actitud o el análisis de evaluación pueden captar mucho mejor la investigación que se ha hecho en el análisis de los sentimientos, incluyendo todas las clases de evaluación (afectivo, cognitivo-conductual) y la relación entre autor, lector y el significado del texto. En base a esta observación y en vista de la teoría de la valoración, en el capítulo 6 se propone un modelo de detección de emociones basado en el conocimiento de sentido común. La definición clara de estos conceptos también ha ayudado a definir de manera adecuada la tarea de análisis de sentimientos en el contexto de los diferentes géneros textuales que empleamos en nuestra investigación. Posteriormente, esta correcta definición ha hecho posible la definición de esquemas de anotación y la creación de recursos para el análisis de los sentimientos en textos de todos los géneros tratados en la investigación. Todos estos recursos fueron empleados tanto en la evaluación de los métodos específicos que hemos creado para el análisis de los sentimientos de los diferentes géneros textuales. Tanto los recursos creados, a través de la alta concordancia entre anotadores obtenida, así como los métodos propuestos, a través del desempeño de los sistemas de aplicación, demostraron que nuestros esfuerzos para dar una definición clara eran de hecho una necesaria contribución a este campo.

2. *¿El análisis de sentimiento puede ser realizado usando los mismos métodos para todos los tipos de texto? ¿Cuáles son las peculiaridades de los diferentes tipos de textos y cómo influyen en los métodos que se utilizan para hacerle frente? ¿Necesitamos recursos especiales para diferentes tipos de textos?*
3. *¿Puede utilizarse un recurso en un idioma dado para resolver problemas en textos escritos en otros idiomas (a través de la traducción)? ¿Cómo pueden ampliarse los recursos a otros idiomas?*

En el capítulo 4, mostramos las particularidades de los diferentes tipos de textos (reseñas, artículos de periódicos, blogs, debates políticos), se analizaron y propusieron técnicas adecuadas para analizar los sentimientos contenidos en cada uno de ellos. Los trabajos presentados en este capítulo fueron publicados en varias conferencias y revistas (incluidas en la sección de referencias bibliográficas y en el listado de contribuciones científicas del Anexo A). Se evaluó cada una de las aproximaciones utilizadas y se demostró que al menos funcionan al nivel de los sistemas más avanzados y, en algunos casos, incluso superan sus resultados.

En este capítulo, presentamos los diferentes métodos y recursos que hemos construido para la tarea de análisis de sentimientos en varios tipos de textos. Se comenzó por explorar los métodos para afrontar la tarea de la minería de opiniones y su resumen basada en rasgos, aplicado en reseñas de productos. Se analizaron las características de esta tarea y se identificaron las carencias en la investigación existente. Propusimos y evaluamos diferentes métodos para superar los obstáculos identificados, entre los cuales los más importantes fueron el descubrimiento de las características mencionadas y de computación indirecta de la polaridad de opiniones de una manera que es una característica dependiente, con la Distancia Normalizada de Google (Cilibrasi and Vitanyi, 2006) y el Análisis Semántico Latente (Deerwester et al., 1990) como medida de asociación semantica. Con posterioridad, se propuso un modelo unificado para la anotación del sentimiento en este tipo de textos (Balahur and Montoyo, 2009), capaz de capturar los fenómenos importantes que hemos identificado – los diferentes tipos de expresiones de sentimiento: directos, indirectos, implícitos; la característica que citan y el fragmento de texto que expresa una opinión específica. Esta distinción no se había propuesto hasta el momento en la literatura. La contribución no sólo se da por el proceso de anotación, sino también por el hecho de que la consideración de grandes tramos de texto en la representación de una sola opinión nos ha llevado a la investigación en búsqueda de respuestas para preguntas de opinión (véase el capítulo 5). La recuperación de texto utilizando fragmentos de tres oraciones puede mejorar de forma considerable el rendimiento del sistema de búsqueda de respuesta en estas preguntas. Conjuntamente con este esquema de anotación, se propuso un método para detectar y clasificar la opinión que aparece sobre las características más importantes de un producto, asignando un número de estrellas en base, utilizando como base un sistema de implicación textual. Este método permite obtener, además de una clasificación de dos vías de opinión sobre las características del producto, un resumen de las frases más importantes que hacen referencia a ellos. Este resumen puede ser empleado como apoyo en la tarea de minería de opiniones basado en funciones y en el proceso de compresión, que ofrece un resumen basado en las opiniones expresadas acerca del producto en cuestión.

Más adelante, se exploraron diferentes métodos para abordar el análisis de los sentimientos de artículos periodísticos. Tras los experimentos iniciales, se analizaron las razones del bajo rendimiento obtenido y se redefinió la tarea, teniendo en cuenta las características de este género textual. Creamos un modelo de anotación y anotamos dos corpus diferentes sobre citas en artículo de prensa, en inglés y alemán. Una vez redefinida la tarea y delimitado el alcance del proceso de análisis de sentimientos a las citas – pequeños fragmentos de texto que contiene el discurso directo, cuyo origen y destino son previamente conocidos, el acuerdo de anotación aumentó significativamente. Asimismo, la redefinición de la tarea hizo posible la aplicación de métodos de procesamiento automático más apropiados que consiguieron mejorar significativamente el rendimiento del sistema de análisis de sentimientos que se había diseñado.

En cuanto a la aplicación del análisis de sentimientos a los diferentes tipos de textos, que contienen una mezcla de información afectiva con información factual y donde las fuentes y los objetivos de las opiniones son múltiples, hemos propuesto diferentes métodos generales para solventar la tarea, en este caso aplicado sobre textos con debates políticos. Los resultados de este último experimento nos motivaron para analizar los requisitos de un esquema general de etiquetado para la tarea de análisis de sentimientos, que pudiera ser usado para capturar todos los fenómenos relevantes en la expresión del sentimiento.

Para ello, en Boldrini et al. (2009), definimos EmotiBlog, un esquema de anotación que permite capturar, con una alta granularidad, todos los fenómenos lingüísticos relacionados con la expresión sentimiento en el texto. Los experimentos posteriores han demostrado que este modelo es apropiado para la formación de modelos de aprendizaje automático para esta tarea en textos de diferentes géneros en los dos idiomas en los que se han realizado los experimentos – inglés y español.

Por último, hemos demostrado que el corpus anotado con EmotiBlog se puede utilizar para extraer las características de los modelos de aprendizaje automático que permitan abordar el análisis de los sentimientos en otros idiomas, a través de un proceso de traducción. Los buenos resultados obtenidos en la tarea 18 de SemEval 2010 – Desambiguación del sentimiento ambiguo competencia adjetivos, donde tradujimos textos en inglés a chino tradicional y aplicamos un modelo de aprendizaje automático entrenado con los datos de EmotiBlog en inglés, demostraron que el modelo de anotación suficientemente robusto y útil incluso con datos que contengan ruido.

4. *¿Cómo podemos tratar las opiniones en el contexto de las tareas tradicionales del PLN? ¿Cómo podemos adaptar las tareas tradicionales (Recuperación de Información, de Respuestas, resumen de texto) en el contexto de los textos que contienen opiniones? ¿Cuáles son los "nuevos" desafíos en este contexto?*

En el capítulo 4, únicamente nos concentramos en la tarea de análisis de sentimientos como un reto independiente, omitiendo los pasos necesarios para obtener los textos sobre los que se aplican los métodos de análisis de opinión o para eliminar la redundancia en la información obtenida. En un escenario del mundo real, sin embargo, la detección automática de la opinión expresada en un texto a menudo no es la primera ni la última tarea por realizar. Previo al análisis de los sentimientos contenidos en los textos, las opiniones sobre ciertos objetivos deberían ser recuperadas. En un gran número de casos, los resultados obtenidos tras determinar automáticamente el sentimiento contenido en los textos siguen planteando muchos problemas en términos de volumen. Así, aunque se extrae el sentimiento de forma automática, todavía puede ser necesario un componente de compresión, con el fin de reducir aún más la cantidad de información, de modo que se puede ser leída y utilizada por una persona.

Teniendo en cuenta estas necesidades, en el capítulo 5 se investigó sobre los métodos para combinar la minería de opiniones con búsqueda de respuestas y generación de resúmenes automáticos. En este capítulo, demostramos que la realización de las tareas tradicionales en el contexto del texto que contiene opiniones tiene muchos retos y que los sistemas que fueron diseñados para trabajar exclusivamente con datos factuales no son capaces de contestar a preguntas de opinión. Nuestra contribución reside en la demostración, mediante un proceso de evaluación, que para poder tratar consultas de este tipo, nuevos elementos tienen que ser definidos. Por ello, propusimos la inclusión de los siguientes elementos: tipo de polaridad, fuente previsto, objeto de la opinión previsto; y definimos los métodos para detectar la fuente y el objeto de los sentimientos expresados, así como las respuestas candidatas, utilizando técnicas de minería de opiniones y el empleo de anotaciones semánticas. Para el proceso de recuperación, como hemos demostrado antes y confirmado en el escenario Búsqueda de Respuestas en Opiniones (en inglés Opinion Question Answering – OQA), utilizar grandes fragmentos de textos es más apropiado cuando se trata con contenido que incluye expresiones de emoción. En concreto, demostramos que la recuperación de fragmentos de tres oraciones conduce a mejores resultados en el caso de los sistemas de OQA. Al proponer y evaluar estos nuevos elementos y técnicas, que han mostrado la forma de abordar OQA de una manera sólida, coherente con las características de los textos con opiniones. Por último, se evaluó el impacto del uso de diferentes herramientas y recursos en esta tarea, como las para la resolución de la anáfora y la ampliación de la pregunta por medio de paráfrasis. Estudiar la forma óptima en que estos dos componentes son combinados en un sistema es una importante línea de trabajo futuro abierta.

En el caso de los resúmenes de opinión, nuestra aportación reside en: a) estudiar el orden en que los sistemas de minería de opiniones y de compresión tienen que ser empleados, b) estudiar la manera en que los sistemas de minería de opinión y de resumen automático puede ser utilizados en conjunto, c) estudiar el efecto de la detección del tema para la minería de opiniones en el contexto de resúmen de opinión; d) proponer un método para resumir opiniones basado en la intensidad del sentimiento expresado.

En esta parte de la investigación, nuestra aportación reside en mostrar que el sistema de análisis de sentimientos debe ser empleado antes en el sistema de compresión y que el componente de análisis de sentimientos debe ser mejorado con mecanismos de detección del tema. En otros casos, aunque los sistemas de análisis de emociones realicen una clasificación correcta (de acuerdo a la polaridad del sentimiento), el hecho de que la relación con el tema no está contemplada conduce a la introducción de los datos no relevantes en los resúmenes finales. Por último, demostramos que en el caso del texto que contiene expresiones de opinión, la relevancia está dada no sólo por la información contenida, pero también por la polaridad de la opinión y su intensidad. Aunque los resultados iniciales han demostrado que no existe una correlación entre las anotaciones del estándar de anotación y el nivel de intensidad de los sentimientos detectados, el sistema de análisis de sentimientos que usó este método en la competición TAC 2008, obtuvo resultados altos en cuanto a la medida-F. Por tanto, creemos que se deben estudiar otros mecanismos para la medición de la intensidad de opinión, de modo que sea posible descubrir la conexión entre la importancia del contenido con opinión en cuanto a la información que aporta, y la intensidad que tiene.

5. *¿Podemos proponer un modelo para detectar las emociones de un texto, en los casos en que se expresa de forma implícita, que requieren conocimiento del mundo para su detección?*

En los primeros capítulos de esta tesis, exploramos la tarea de análisis de sentimientos en diferentes tipos de textos e idiomas y proponemos una amplia variedad de métodos apropiados para afrontar los problemas que aparecieron para cada tipo de texto. La mayoría de las veces, sin embargo, los diferentes enfoques se han limitado a conocer sólo las situaciones donde el sentimiento se expresó de forma explícita, es decir, cuando se pueden encontrar en el texto señales lingüísticas que indiquen que contiene elementos subjetivos o sentimiento. No obstante, en muchos casos, la emoción en la que se basa el sentimiento no está explícitamente presente en el texto, pero se puede inferir en base al conocimiento de sentido común (es decir, la emoción no está explícitamente, pero implícitamente expresada por el autor, mediante la presentación de situaciones que la mayoría de las personas, gracias a su sentido común, asocian con una emoción, como "ir a una fiesta", "ver a su hijo dar sus primeros pasos", etc).

En el capítulo 6 de la tesis, presentamos nuestra contribución a la cuestión de la detección automática de emoción expresada en el texto de manera implícita. La aproximación inicial, sustentada sobre la Teoría de la Relevancia (Sperber and Wilson, 2000), se basa en la idea de que la emoción es provocada por conceptos específicos, de acuerdo a su importancia, y debe considerarse en relación con las necesidades y motivaciones básicas.

El segundo enfoque que proponemos se basa en la Teoría de los modelos de evaluación (presentada en detalle por Scherer, 1999). La idea general que subyace es que la mayoría de veces las emociones no se declaran explícitamente en los textos, sino que resultan de la interpretación (evaluación) de las acciones contenidas en la situación descrita, así como las propiedades de sus actores y objetos. Nuestra contribución en esta última parte de la investigación reside en la creación de un marco para la representación de situaciones que se describen en el texto como cadenas de acción (con sus correspondientes actores y objetos), y sus propiedades correspondientes (incluyendo las afectivas), como el conocimiento de sentido común. Se muestra la forma de detectar automáticamente a partir del texto los llamados "criterios de evaluación" y la forma de extender el conocimiento sobre las propiedades de los conceptos involucrados en cada situación utilizando fuentes externas. Por último, se demuestra a través de una extensa evaluación de que tal representación es útil para obtener una etiqueta exacta de la emoción expresada en el texto, sin ninguna pista lingüística está presente en él. Debido a la relación directa entre los sentimientos y la presencia de emociones, la detección de expresiones implícitas de emoción puede aumentar el rendimiento de los sistemas de análisis de sentimientos, un hecho que hemos demostrado en los experimentos que presentamos en el capítulo 3 en el caso de análisis de sentimientos de comentarios.

CONCLUSIONES Y TRABAJOS FUTUROS

Finalmente, se presentan las posibles direcciones de desarrollo futuro de la investigación realizada en esta tesis. Con el fin de ser coherente con la estructura de la tesis, estas direcciones se presentan de la misma forma, en relación con las preguntas de investigación que nos propusimos a responder y los puntos que siguen pendientes de resolución:

i. *¿Cómo se puede definir la tarea de análisis de sentimientos y, en una perspectiva más amplia, la minería de opiniones, de una manera correcta? ¿Cuáles son los conceptos principales que se deberían definir antes de afrontar la tarea?*

En el capítulo 2 de la tesis, presentamos un resumen de las definiciones que se dan en la literatura de PLN para las tareas relacionadas con el análisis de la subjetividad, el análisis de los sentimientos, la minería opinión, evaluación/análisis de la actitud, la detección de las emociones, así como los conceptos que participan en dichas tareas. Demostramos posteriormente que hay una alta inconsistencia entre las definiciones de las tareas. Finalmente, propusimos un conjunto de definiciones operativas coherentes con la manera en que los diferentes términos relacionados con el análisis de los sentimientos se definieron en fuentes bien establecidas y los planteamientos que hicimos en la investigación. En el capítulo 4, mostramos que el análisis de los sentimientos debe de ser abordado de una manera diferente, dependiendo de los tipos de texto considerado. A partir de estos esfuerzos, una futura línea de trabajo es la definición de un marco unificado para el análisis de los sentimientos, que describa la tarea de una manera general, aunque uniforme a través de géneros y aplicaciones. En este sentido, demostramos en los experimentos con artículos de prensa, dicho marco debe ser construido tomando en cuenta no sólo el contenido textual, sino también los elementos que se relacionan con el autor y el lector.

Las siguientes preguntas de investigación abordadas en esta tesis fueron:

2. *¿El análisis de sentimiento puede ser realizado usando los mismos métodos para todos los tipos de texto? ¿Cuáles son las peculiaridades de los*

diferentes tipos de textos y cómo influyen en los métodos que se utilizan para hacerle frente? ¿Necesitamos recursos especiales para diferentes tipos de textos?

3. *¿Puede utilizarse un recurso en un idioma dado para resolver problemas en textos escritos en otros idiomas (a través de la traducción)? ¿Cómo pueden ampliarse los recursos a otros idiomas?*

En el Capítulo 4, presentamos los diferentes métodos y recursos para la tarea de análisis de sentimientos en diferentes tipos de textos (reseñas, artículos de periódicos, blogs y debates políticos) en diferentes idiomas (inglés, español y alemán). Una futura línea de trabajo es la extensión de los recursos propuestos para otros idiomas, ya sea a través de la traducción, en cuyo caso, los recursos deberían ser refinados y evaluados para los diferentes tipos de textos en su idioma original, o por anotación directa en el idioma de destino. En este último caso, sería interesante estudiar las diferentes formas de expresar los sentimientos en dependencia de la cultura y el idioma de los textos. En relación directa con el trabajo desarrollado en esta tesis, en diferentes tipos de textos e idiomas, las futuras líneas de investigación podrían ser:

i. En el caso de los textos de revisiones:

   a) La extracción automática de taxonomías de las características del producto.

   b. La extensión del marco de trabajo para la anotación de revisiones y la minería de opiniones basada en rasgos y resúmenes en otros idiomas diferentes del inglés y español.

ii. En el contexto de las citas de periódicos y, de manera más general, artículos de prensa:

   a) El estudio del impacto de la fuente de noticias (es decir, en términos de sesgo, la reputación, la confianza) en el proceso de análisis de sentimientos, dentro de las fuentes del mismo país/cultura y entre países y culturas diferentes;

   b) El estudio de la influencia del lector sobre la forma en que el sentimiento es percibido de los artículos periodísticos;

   c) La extensión automática y semiautomática de los recursos que construidos a otros idiomas (es decir, en adición a la colección de citas se anotaron para inglés y alemán);

   d) El desarrollo de un marco para el análisis de los sentimientos que se tienen en cuenta los tres componentes propuestos en el texto – autor, lector y texto - y la manera en que interactúan en la negociación el significado del texto, de acuerdo con el acto de habla y las teorías de evaluación;

  - e) El estudio de la influencia del contenido de las noticias (es decir, lo que denota como "una buena noticia frente a lo malo") en la forma en que el sentimiento se expresa y, posteriormente, en el desempeño de la tarea sentimiento análisis automático.

- iii. En el contexto de los debates políticos y textos de carácter general:
  - a) Estudio de los métodos para representar la estructura de diálogo y la influencia de los diferentes métodos de análisis del discurso, así como las herramientas (por ejemplo, para la resolución de la anáfora) sobre los métodos de análisis propuestos sentimiento;
  - b) El estudio y el uso de técnicas de modelado tema a fin de detectar con precisión el objetivo del sentimiento expresado, independientemente de su naturaleza que se conoce de antemano;
  - c) El estudio de cómo afectan las técnicas de argumentación con el fin de detectar el uso de disparadores de emoción en relación con el tema de discusión.
- iv. En el contexto de los blogs:
  - a) La extensión del método propuesto de anotación y etiquetado de corpus a otros idiomas y tipos similares de texto (por ejemplo, foros, paneles de discusión).
  - b) El estudio y el uso de técnicas de modelado del tema, así como de los efectos de la resolución co-referencia en un nivel intertextual.

4. *¿Cómo podemos tratar las opiniones en el contexto de las tareas tradicionales del PLN? ¿Cómo podemos adaptar las tareas tradicionales (Recuperación de Información, de Respuestas, texto de resumen) en el contexto del contenido que expresa opinión? ¿Cuáles son los "nuevos" desafíos en este contexto?*

Teniendo en cuenta estas necesidades de aplicaciones del mundo real, que necesitan, aparte del componente de análisis de sentimientos, también la recuperación de texto y los componentes de texto de resumen, en el capítulo 5 se propone métodos para combinar la minería opinión con búsqueda de respuestas y generación de resúmenes automáticos. Demostramos que la realización de las tareas tradicionales en el contexto del texto de tipo opinión tiene muchos retos y que los sistemas que fueron diseñados para trabajar exclusivamente con datos factuales no son capaces de contestar a preguntas de opinión. En esta tesis se propone nuevos métodos y técnicas para adaptarse y responder la pregunta de los sistemas de compresión para tratar el contenido dogmático.

En el caso de los sistemas de búsqueda de respuestas a preguntas de opinión, el trabajo futuro incluye el desarrollo de un punto de referencia para la clasificación de preguntas de opinión y la propuesta de métodos adecuados para hacer frente a cada tipo de consultas de opinión, en un entorno monolingüe, multilingüe y entre idiomas. Además, el marco para responder a preguntas de opinión debe ampliarse con los recursos adecuados a otros idiomas. Como apreciamos en nuestros experimentos en el concurso MOAT NTCIR 8, existe una necesidad inmediata para incluir métodos de alto rendimiento para la resolución de expresiones temporales y la resolución de la anáfora. Lamentablemente, debido al bajo rendimiento de los sistemas de resolución de estos aspectos, la influencia que tienen sobre los sistemas de búsqueda de respuestas a preguntas de opinión es negativa. Otra línea de trabajo futuro es el estudio de las técnicas de ampliación de consultas que sean apropiadas para el contenido de tipo opinión. Por lo visto en nuestros experimentos, el uso de una colección de paráfrasis que no está específicamente diseñado para el contenido textual que contiene sentimientos conduce a una caída en el rendimiento del sistema final. En el caso de los enfoques resumen automático de texto con sentimientos, las líneas de trabajo futuro incluyen el desarrollo de técnicas adecuadas para la detección de la relevancia y el estudio de medidas cualitativas para la evaluación de los métodos de compresión de opinión.

La última pregunta de investigación que aborda en esta tesis fue:

5. *¿Podemos proponer un modelo para detectar las emociones de un texto, en los casos en que se expresa de forma implícita, que requieren conocimiento del mundo para su detección?*

En el capítulo 6 de la tesis, propusimos dos métodos para detectar expresiones implícitas de la emoción, de situaciones donde la etiqueta afectiva puede inferirse basada en el conocimiento de sentido común (es decir, la emoción no está explícitamente, sino implícitamente expresada por el autor, mediante la presentación de situaciones que la mayoría de la gente, basándose en el sentido común, sus necesidades y motivaciones, asocian con una emoción, como "guerra", "terrorismo", "ir a una fiesta", "ver a su hijo dar sus primeros pasos", etc.). El planteamiento inicial que se propone se basa en la idea de que la emoción es provocada por conceptos específicos, de acuerdo a su tema, se ve en relación con las necesidades y motivaciones básicas. Posteriormente, se propone un marco para la detección de emoción en texto basado en la teoría de la valoración (Appraisal Theory), el modelado de la posible interpretación (evaluación) de las acciones contenidas en la situación descrita, de acuerdo con el conocimiento de sentido común sobre las acciones y las propiedades de sus actores y objetos. El trabajo futuro incluye la ampliación de los recursos creados para otros idiomas y la ampliación de la base de conocimiento de sentido común de situaciones que puedan suscitar una emoción, con fuentes externas de conocimiento. En este contexto, las líneas de interés para la labor futura podrían ser:

- Las técnicas de desarrollo para la personalización y adaptación del contenido de la información en función del contexto de usuario, a través de la explotación de los contenidos subjetivos generados por el usuario, así como el análisis de las preferencias expresadas de forma explícita, a través de opciones en los perfiles de usuario y de forma implícita, a través de las conexiones en las redes sociales, las contribuciones a los foros, comentarios, blogs o microblogs;
- El diseño, la implementación, la población y la ampliación de una base de conocimiento para el modelado de preferencias de usuarios, para poder realizar minería de opiniones de forma personalizada (es decir, en función de los objetivos del usuario, las creencias, opiniones, observaciones, sentimientos). En este contexto, las capacidades de búsqueda semántica podrían ser mejoradas y ampliadas de una manera más centrada en el usuario.

Reunido el Tribunal que suscribe en el día de la fecha acordó otorgar, por a la Tesis Doctoral de Don/Dña. la calificación de .

Alicante de de

El Secretario,

El Presidente,

UNIVERSIDAD DE ALICANTE
CEDIP

La presente Tesis de D. ____ ha sido registrada con el nº ____________ del registro de entrada correspondiente.

Alicante ___ de __________ de _____

El Encargado del Registro,

La defensa de la tesis doctoral realizada por D/Dª se ha realizado en las siguientes lenguas: y , lo que unido al cumplimiento del resto de requisitos establecidos en la Normativa propia de la UA le otorga la mención de “Doctor Europeo”.

Alicante, de de

EL SECRETARIO EL PRESIDENTE