Journal of Intelligent Information Systems manuscript No.
(will be inserted by the editor)

Towards a Unified Framework for Opinion Retrieval, Mining and Summarization

Elena Lloret · Alexandra Balahur · José M. Gómez · Andrés Montoyo · Manuel Palomar

Received: disate / Accepted: date

Abstract The exponential increase of subjective, user-generated content since the birth of the Social Web, has led to the necessity of developing automatic text processing systems able to extract, process and present relevant knowledge. In this paper, we tackle the Opinion Retrieval, Mining and Summarization task, by proposing a unified framework, composed of three crucial components (information retrieval, opinion mining and text summarization) that allow the retrieval, classification and summarization of subjective information. An extensive analysis is conducted, where different configurations of the framework are suggested and analyzed, in order to determine which is the best one, and under which conditions. The evaluation carried out and the results obtained shows the appropriateness of the individual components, as well as the framework as a whole. By achieving an improvement over 10% compared to the state-of-the-art approaches in the context of blogs, we can conclude that subjective text can be efficiently dealt with by means of our proposed framework.

Keywords Intelligent System · Opinion Retrieval, Mining and Summarization Framework · Information Retrieval · Opinion Mining · Text Summarization

1 Introduction

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

Elena Lloret · Alexandra Balahur · José M. Gómez · Andrés Montoyo · Manuel Palomar
Department of Software and Computing Systems
University of Alicante
Apdo. de correos, 99, E-03080 Alicante.
Tel.: +34-96-5903772
Fax: +34-96-5909326
E-mail: {elloret,abalahur,jmgomez,montoyo,mpalomar}@dlsi.ua.es

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 in events, by expressing their opinions on them, by commenting on the news appearing and the events that take place in all spheres of 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 behaviour 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. For instance, between 73% and 87% of Internet users have reported that the opinions expressed on the Internet had a significant influence on their decisions (Pang and Lee, 2008). What people think of a product, service, etc. is of great value for companies, because they can monitor public perception on their products, services, policies, etc. which later can be used to strengthen their competitivity in the market. Given the proven importance of subjective, user-generated content and the fact that it is increasing at an exponential rate, automatic text processing systems must be built to extract the relevant knowledge from such user-generated content.

The treatment of subjective information has become one of the most dynamic research fields in Natural Language Processing (NLP), within the tasks of subjectivity analysis and of opinion mining (also called sentiment analysis). Opinion Mining (OM) can be briefly defined as the task of determining the sentiment and attitude of a source with respect to some "target". The term "sentiment" represents a settled opinion reflective of ones 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, 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).

Apart from this NLP research area, which treats subjective information, the vast amount of sources from which this information is gathered and its inherent redundancy makes it impossible to manage, thus requiring systems and tools able to provide efficient mechanisms to face the problem of the information overload. In light of this, other NLP research areas can be considered suitable to provide mechanisms that allow users to manage the information efficiently, reducing to a great extent the time a user would need to process it on its own. On the one hand, Information Retrieval (IR) reduces the search space for users, selecting the documents that may be relevant according to a query. In this manner, users are able to find specific information quickly. On the other hand, Text Summarization (TS) provides a condensed version of one or several documents (i.e., a summary) which can be used as a substitute of the original texts. Such summaries can be tailored to specific information depending on user interests.

Many approaches and systems have been developed to tackle each of the aforementioned problems separately, or combining two specific fields together within the same approach (e.g., Swotti[1] is able to retrieve and analysis the opinions of a document). Through the literature, when different NLP research areas have been combined, it has been proven that the overall performance of the resulting approach increased, besides extending its capabilities (Stoyanov and Cardie, 2006). For instance, approaches found in (Yang and Liu, 2008), (Torres-Moreno et al, 2009) or (Jin et al, 2009) integrate TS with IR, question answering and OM, respectively, showing the benefits of the proposed system combinations with respect to using them alone. However, to the best of our knowledge, very little research has been carried out into an in depth analysis of the benefits and limitations of a fully automatically approach that combines IR, OM, and TS.

One of the main reasons might be the difficulty of the task itself. Being able to combine these three applications together with the aim of outputting a coherent text fragment is quite an ambitous goal. In 2008, an *Opinion Summarization Pilot* task, was set up by the first time within the *Text Analysis Conference*[2] (TAC). The task consisted of generating fluent and coherent summaries from opinion questions whose answer could be found in blogs. In order to simplify the task, the organizers provided a list of short sentences which already contained the answers to the questions. Despite this fact, due to the difficulty associated with the task, it was substituted for other types of summarization tasks in the next years (Dang and Owczarzak, 2009).

Another important issue to bear in mind is the moderate performance current TS systems (around 45% and 30% for single- and multi-document summarization, respectively for the F-measure), thus hypothesizing about the fact that the performance of the approach would decrease significantly. However, several extrinsic evaluations conducted in TS have been proven, that although imperfect in their nature, summaries can be beneficial when combined with other NLP tasks (Mani et al, 1999), (Shen et al, 2007), (Lloret et al, 2010).

Therefore, the objective of this paper is to present a unified framework, fully automatic, which integrates IR, OM and TS together. The concept "unified" is used to denote the fact that we aim at developing a single and integrated process, where the output of the IR component is the input for the OM, and the output of the OM will be the input for TS. In this manner, with the proposed framework, we can automatically retrieve subjective content from the Web, analyze and classify the opinions found, and finally produce a summary, that will contain the specific information a user is looking for. Consequently, the main contribution of this paper is to explore the task of opinion retrieval, mining and summarization through the proposal of the aforementioned unified

[1] www.swotti.com

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

framework, which is novel in the context of the opinion analysis. Furthermore, the analysis of different IR and OM approaches, as well as different compression rates for summaries is also provided. The results obtained on blogs show that the proposed approach is very competitive and encouraging despite the difficulty of the task.

This paper is structured as follows. Section 2 gives an overview on previous work on different approaches of IR, OM and TS, as well as the existing approaches combining them. In Section 3, the different approaches used for IR, OM and TS are explained in detailed. Section 4 describes the extensive set of experiments carried out. The results obtained together with a detailed analysis of the benefits and limitations of the proposed approached are discussed in Section 5. Finally, some relevant conclusions are drawn and further work is explained in Section 6.

2 Related Work

This section aims at providing the state of the art of the different NLP areas involved in this research work (i.e., IR, OM and TS), and the possible existing combinations. Therefore, a brief presentation of the most significant challenges and proposed approaches in IR, OM, and TS is first given in Sections 2.1, 2.2 and 2.3, respectively. In this manner, the definition and aims of each research field, as well as the common approaches to tackle them, are provided.

Further on, Sections 2.4, 2.5 and 2.6 focus on the different combinations that have been proposed, where IR, OM, and TS have been combined into a single approach (i.e., OM with IR; OM with TS; or IR with TS). To the best of our knowledge these combinations have been taken place two at a time, and no previous attempt to build a single process where IR, OM and TS are integrated within the same framework has been approached so far. For clarity purposes, Table 1 shows the combinations that we are going to explain in this section.

Information retrieval	✓
Opinion Mining	✓
Text Summarizaiton	✓
Information retrieval & Opinion Mining	✓
Opinion Mining & Text Summarization	✓
Information retrieval & Text Summarization	✓
Information retrieval & Opinion Mining & Text Summarization	✗

Table 1 Outline of the state-of-the-art combinations for Information Retrieval, Opinion Mining and Text Summarization.

2.1 Information Retrieval

Information Retrieval (IR) is the discipline that deals with the retrieval of information (mostly documents) in response to a query or topic statement. The need for effective methods to automate IR has become acute because of the tremendous explosion in the amount of information, and the very high and growing number of document sources on the Internet. IR techniques have been applied to different types of documents, such as images, audio, video or geographic data (Datta et al, 2008), (Aslandogan and Yu, 1999), (Foote, 1999). Nonetheless, research has especially focused on the retrieval of text, and more specifically, on natural language text. Due to the size of the Internet, there is a need to use structures that minimize temporal and spatial costs. The most common structure is inverted indexes (Witten et al, 1999). To build these inverted indexes it is necessary to pre-process every document in order to extract all its terms. Each term of the collection has a pointer to the documents it appears in, so the searching for documents containing the specific terms is instantaneous. When indexing the Internet, there is a necessary tool: the web crawler (Najork and Heydon, 2002). The objective of crawling is to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them, in a robust and polite way (Manning et al, 2008). The link structure is analyzed to extract the most important URLs, and this data is used to rank the results and give download priority to them at indexing time. A well-known algorithm of link analysis is PageRank (Brin and Page, 1998).

2.2 Opinion Mining

Opinion Mining is the task of analyzing, classifying and extracting the subjective information and the sentiment associated to a specific target. From the computational perspective, it requires different techniques, 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 opinions 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. 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. It is also important to note the fact that opinion mining does not necessarily require as input an opinionated piece of text (Pang and

Lee, 2008). For instance, in Koppel and Shtrimberg (2006), good versus bad news classification has also been considered as a sentiment classification task.

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) 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.

The issue of extracting and classifying opinions from text has been tackled at different text levels: document level, sentence level and feature level. Research in document-level sentiment classification includes work by Turney (2002), Pang et al. (2002), Dave et al. (2003), Pang and Lee (2003), Chaovalit and Zhou (2005), Ng et al. (2006), or Kudo and Matsumoto (2004). 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), (Riloff and Wiebe, 2003), (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. Finally, sentiment analysis at the feature level, also known as "feature-based opinion mining" ((Hu and Liu, 2004) and (Liu, 2006)), 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).

2.3 Text Summarization

Text Summarization identifies the most important information in a document or documents and extracts it in the form of a summary. Many approaches have been developed for automatically determining which information has to be included in the summary. These include statistical techniques such as term-frequency and inverse document frequency (McCargar, 2005); textual entailment (Lloret et al, 2008a); topic signatures (Lin and Hovy, 2000); topic identification (Harabagiu and Lacatusu, 2005), informativeness and event words (Kuo and Chen, 2008), or the use of graph-based algorithms (Giannakopoulos et al, 2008). Recently, the development of the Internet, the availability of massive textual databases together with international evaluation efforts, such as DUC[3] and TAC, have fuelled research in this field. TS research has evolved together with the user needs and the manner in which they express their needs on the Internet. For example, TS approaches that produce query-focused (Ma et al, 2008), (Zhao et al, 2009) and opinion-oriented summaries (Carenini and Cheung, 2008), (Lloret et al, 2009), (Balahur et al, 2010c), (Balahur et al, 2009c), (Kabadjov et al, 2009) have become more important in the last years. Moreover, although most of the approaches still rely on a sentence-extraction paradigm where several features are employed to determine the importance of sentences in documents and then select and extract the most relevant ones to build the summary (Mani, 2001), the improvement of natural language generation and sentence fusion and simplification methods are encouraging approaches to generate summaries following an abstractive strategy. Examples of these types of summaries can be found in (Ou et al, 2007), (Sauper and Barzilay, 2009), and (Saggion, 2009) to name a few. Another issue that is changing in the TS field is the domain of the documents for generating the summaries. Traditionally, newswire and scientific articles have been the most common domains to perform TS on (Hsin-Hsi and Chuan-Jie, 2000), (McKeown and Radev, 1999), (Teufel and Moens, 2002), but currently, a wide range of novel domains has been exploited as well, including legal domain (Cesarano et al, 2007), short stories (Kazantseva, 2006), books (Mihalcea and Ceylan, 2007), or image captioning (Plaza et al, 2010).

[3] Document Understanding Conference http://duc.nist.gov (*Last Access: 06/02/2012*)

2.4 Information Retrieval with Opinion Mining

Information retrieval can be combined with opinion mining, thus helping to deal with subjective information instead of factual one. This combination is known as opinion retrieval. More concretely, opinion retrieval is the process of searching and providing documents related to an opinion expressed to a topic, product, entity, etc.. It requires documents to be retrieved and ranked according to the opinions they have about a query topic. A relevant document must contain opinions about the query (positive or negative). This new kind of retrieval has become more relevant because of the rising of the interactivity on the Internet: users can express their opinions about different topics, but they also want to have access to other users' opinions.

The major part of the research on this area can be found in the TREC Blog track (Ounis et al, 2006) where IR and OM systems are used together in order to find relevant opinionated documents to a query. The approaches presented focused on detecting the subjectivity for each document, using different OM techniques such as opinion expression dictionaries (Mishne, 2006), machine learning algorithms (Zhang et al, 2007) or proximity and phrase matching (Yang, 2008). Then, IR techniques are applied in order to provide first the highest relevant documents from the whole set of the opinionated documents found.

In the business context, we can find several opinion retrieval systems such as Ciao[4] or Swotti[5], which extract opinions of a particular product using different techniques. Ciao has a database with opinions for different products, which are accessible from a standard product search engine, selecting the desired product. In Ciao, the users explicitly insert the opinions for a particular product. Swotti is closer to the opinion retrieval approach. While indexing, it extracts the products and their features, searches for opinions of them and classifies them by their polarity. But it does not make a ranking of opinions, it only focuses on the rating assigned to the product features. In neither of them, the techniques used for the implementation are available.

2.5 Opinion Mining with Text Summarization

The combination of opinion mining and text summarization has resulted in the specific task of opinion summarization, whose aim is the generation of opinion-oriented summaries. This type of summaries take into consideration the sentiment a person has towards a topic, product, place, service, etc. Opinion Mining provides the sentiment associated with a document at different levels (document, fragment, sentence or even word-level), and text summarization will identify the most relevant parts of a document and produces from them a coherent fragment of text (the summary). Therefore, opinions have to be first detected and classified according to the orientation of the sentiment

[4] http://www.ciao.co.uk/ (*Last Access: 06/02/2012*)

[5] http://www.swotti.com (*Last Access: 06/02/2012*)

they express (i.e. their polarity - positive, negative or neutral). Further on, TS will be in charge of determining the sentences to be included in the summary.

This new task is motivated by the fact that in the recent years, the subjectivity appearing in documents, in addition to the possibility users have for expressing whatever they like on the Internet, has led to new emerging textual genres, such as blogs, forums, reviews, wikis, tweets, that have to be treated differently from the traditional objective information. As a consequence, opinion mining techniques are crucial for the correct generation of opinion-oriented summaries.

As previously mentioned in Section 1, the *Opinion Summarization Pilot* task proposed at TAC 2008[6] provided a very suitable evaluation framework to test different Opinion Question Answering, OM and TS approaches in tandem. Most participants employed techniques based on the already existing summarization systems, but adding new features (sentiment) to detect and classify the opinions. That was the case of CLASSY (Conroy and Schlesinger, 2008), LIPN (Bossard et al, 2008), and CCNU (He et al, 2008) systems. Other approaches, such as (Balahur et al, 2008) and (Cruz et al, 2008) focused on the retrieval and filtering stage taking the polarity into account.

Out of the scope of the TAC competition, we can find other interesting approaches, as well. For instance, in (Beineke et al, 2003) machine learning algorithms are used to determine which sentences should belong to a summary, after identifying possible opinion text spans. The features found to be useful to locate opinion quotations within a text included location within the paragraph and document, and the type of words they contained. Similarly, in (Zhuang et al, 2006) the relevant feature and opinion words and their polarity (whether a positive sentiment or a negative) are identified, and then, after identifying all valid feature-opinion pairs, a summary is produced, but focusing only in movie reviews. Normally, online reviews also contain numerical ratings that users give when providing a personal opinion about a product or service. The approach described in (Titov and McDonald, 2008) proposed a *Multi-Aspect Sentiment* model. This statistical model uses aspect ratings to discover the corresponding topics and extract fragments of text. In (Lerman and McDonald, 2009), an approach to produce contrastive summaries in the consumer reviews domain is suggested. Contrastive summarization refers to the problem of generating a summary for two entities in order to highlight their differences, for example, different people's sentiments about several products. In order to produce this type of summaries they extend the *Sentiment Aspect Match* model proposed described in (Lerman et al, 2009), originally designed to generate single product opinion summaries.

[6] http://www.nist.gov/tac/2008/summarization/op.summ.08.guidelines.html (*Last Access: 06/02/2012*)

2.6 Infomation Retrieval with Text Summarization

The combination of information retrieval with text summarization helps users to decide whether to read or not a full retrieved document, by just having access to its summary. This allows users to manage the information in a more efficient manner, avoiding them to spend too much time navigating through the documents to see if they are interesting for their needs.

In this sense, the input for the text summarization process is the output of the information retrieval (i.e. the documents retrieved). In the approach proposed in (Kan and Klavans, 2002), summaries are employed to present an alternative visualization of the documents coming from a standard IR framework. Moreover, the optimal length that a summary should have to be useful for users when using them as output of a search engine is analyzed in (Kaisser et al, 2008), finding that the preferred length for the users depended on the query. However, the most common approach is to combine information retrieval and text summarization in this manner: the documents related to a topic are retrieved first, and then, a summary taking into account these documents is generated. Therefore, an IR system will help to gather only relevant documents to a query, while TS systems will select the most important information from them. Radev and Fan Radev and Fan (2000) proposed an open-domain multi-document summarizer that generates summaries from Web search results. Similarly, SWEeT (Steinberger et al, 2008) relies on a search engine to retrieve relevant documents to a user query from the Web, and then summarization techniques based on Latent Semantic Analysis (LSA) are used to identify and extract the most important sentences from the retrieved documents, using at the same time cosine similarity to avoid redundancy in the final summaries. The QCS system (Dunlavy et al, 2007) also integrates a IR module but, instead of retrieving documents directly from the Internet, it does so from a static document collection. Once the relevant documents have been retrieved, the system clusters them according to their main topic, and finally a summary is produced for each cluster. The TS process is performed in two steps. Firstly, a single summary is generated for each document cluster, and then those extracted summary sentences are taken into account to produce the final summary. The way sentences are selected to become part of the summaries is by using a Hidden Markov Model, computing the probability of a sentence with regard to whether it is a good summary sentence or not.

In the literature, we can also find other approaches that combines information retrieval and text summarization in the opposite way, although this use of summaries is not the common one. In this case, summaries are used to benefit the retrieval process, for example at the indexing stage, improving the time to retrieve the documents and the performance of the IR system. In (Sakai and Sparck-Jones, 2001) it was proven that generic summaries with a compression rate ranging from 10% to 30% were the most appropriate for the indexing stage in IR tasks, concluding that a summary index was as effective as the fulltext index, for precision-oriented search of highly relevant documents.

3 A Unified Framework for Opinion Retrieval, Mining and Summarization

The goal of this Section is to present our proposed unified framework which integrates information retrieval, opinion mining and text summarization together. We would like to stress upon the fact that in this research work, the concept "unified" is used to denote the fact that our objective is the development of a single and integrated process, where the output of the IR component is the input for the OM, and the output of the OM will be the input for TS.

Therefore, through this framework it would be possible to search and retrieve subjective content from the Web (more concretely, for this research, we have experimented with blogs), analyze and classify the subjective information found in them, and as a last step, extract the most relevant information, by generating a summary of a desired size.

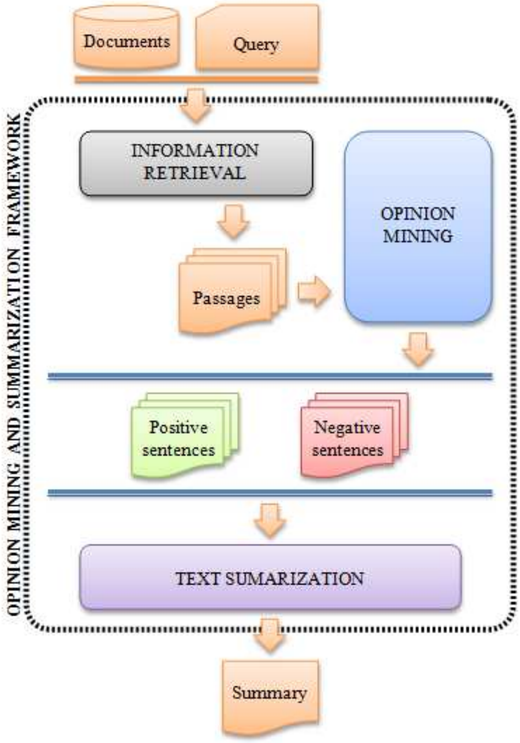

Fig. 1 Our proposed unified framework.

Figure 1 depicts the proposed framework, where the individual components are integrated together.

As it can be seen, the input is a collection of documents (blogs in this specific research) and a query.

The Information Retrieval (IR) module indexes the document collection with the purpose of returning back the results as quick as possible when a user question is received. Such results will consist of a ranked list of passages. A passage is a text fragment from a document and each of these passages contains a similar structures to the user question ones. The size of these passages depend on the system configuration. For our experiments, we used passages composed by 1 or 3 sentences. As an output of the IR module, the passages together with their probability of containing the correct answer are listed. The Opinion Mining (OM) module take these passages and extracts the subjective sentences. To that end, this module uses 4 semantic resources in order to weigh each passage terms. Finally, the OM module obtains the subjectivity of every passage sentence (positive, negative or neutral) using the computed term weights. Moreover, this module proposes another approach which enriches the passages adding new semantic knowledge in the form of related terms by using Latent Semantic Analysis (LSA) techniques. Further on, the Text Summarization (TS) module uses the returned positive and negative sentences from the OM module. With these sentences, this module identifies the most relevant ones using statistic and cognitive-based features after removing the redundant information with textual entailment techniques. The final objective of this process is to generate an non-redundant opinion-oriented extract, containing the most relevant user opinions with regard to the question topic.

The remaining of this section is organized into three subsections, each of them corresponding to the information retrieval (Section 3.1), the opinion mining (Section 3.2) and the text summarization (Section 3.3) components, respectively.

3.1 Information Retrieval

JAVA Information Retrieval system (JIRS) is a IR system especially suitable for question answering tasks. We have chosen this IR system due to the good results that this system achieved in previous international question answering competitions (y Gómez et al, 2005), (Soriano et al, 2005), (Christensen and Ortiz-Arroyo, 2007), (Buscaldi et al, 2010). Its purpose is to find the fragments of text (passages) that are more probable of containing the answer to a user question posed in natural language, instead of just finding the documents that are relevant to a query. To that end, JIRS uses the question structure and tries to find an equal or similar expression in the documents. The more similar the structure between the question and the passage is, the higher the passage relevance. For instance, if the question is "*Why do people like Starbucks better than Dunkin Donuts?*", JIRS will try to find a passage with the expression "*We thought a lot of* ***people like Starbucks better than Dunkin' Donuts****, because its seems nicer or even classier*". In this example, the question and the passage contain the same structure and, in this case, the answer frequently appears close.

JIRS is able to find question structures in a large document collection quickly and efficiently using different n-gram models. To do this, JIRS first uses a traditional passage retrieval system and then searches all possible n-grams of the question in the retrieved passages. Further on, it rates them depending on the number and the weight of the n-grams that appeared in these passages. The system architecture is shown in Figure 2.

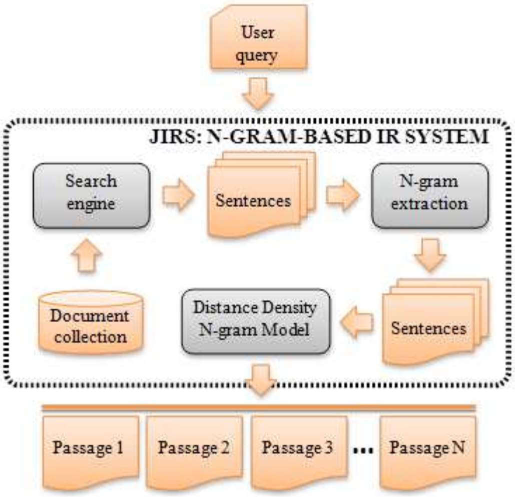

Fig. 2 The main architecture of the JIRS n-gram based PR system

In this figure, we can observe how the user question is passed to the three main modules of JIRS: *search engine*, *n-gram extraction* and *Distance Density N-gram model*. In the first module, the search engine returns a ranked list of m passages with question keywords. The size of this passage depend on the configuration of the system, and it is measured in number of sentences. For each retrieved sentence by the JIRS system, the previous and next sentence from the document is also added, in order to form the final passage. This occurs for passage lengths different from 1. For instance, a passage of length 3 is formed by adding to the retrieved sentence the next and the previous sentences from the document, so that in total the passage will consist in three sentences; a passage length of 5 contains two previous and two next sentences (five sentences in total). In (Gómez, 2007) and (Gómez et al, 2007) it was proven that the highest precision of the system was obtained when passages of 1 or 3 sentences were used.

With each passage that was retrieved by the search engine module, the n-grams formed by the question terms are extracted. Finally, each passage is ranked according to these factors: i) the extracted n-grams; ii) the weight of these n-grams; and iii) the relative distance between the extracted n-grams.

Several n-grams models were compared in (Gómez, 2007) but, for this specific research, we have only used the *Distance Density n-gram* model because it was the best one experimentally obtained. The *Distance Density n-gram* model is based on searching the n-grams with the greatest weight instead of the longest ones. With this model, the final n-gram weight is obtained by multiplying the weight calculated with Formula (1) by a distance factor that takes into account the distance with respect to the n-gram with highest weight.

$$h(x) = \sum_{k=1}^{j} w_k \tag{1}$$

where $w_1, w_2, ..., w_j$ are the term weights of the j-gram $x = t_1 t_2 ... t_j$. These weights should penalize the terms that appear frequently in the document collection (e.g., stopwords) and promote the relevant words (i.e., the question terms that are of crucial importance to retrieve a relevant passage). The following function (Formula 2) was introduced to assign the weight to a term:

$$w_k = 1 - \frac{\log(n_k)}{\log(N+1)} \tag{2}$$

where n_k is the number of passages in which the term t_k appears, and N is the number of system passages. We make the assumption that stopwords occur in every passage (i.e., n_k takes the value of N). For instance, if the term t_k occurs only once in the passage collection, its weight will be equal to 1 (the greatest weight). However, if it is a stopword, its weight will be the lowest one.

Therefore, the similarity value depends on the density of the question terms in the passage, and it is calculated as the sum of all n-gram weights, multiplied by the distance factor and divided by the sum of all term weights of the question. The formula we have used is the following:

$$Sim(p,q) = \frac{1}{\sum\limits_{i=1}^{n} w_i} \cdot \sum_{\forall x \in \hat{P}} h(x) \frac{1}{d(x, x_{max})} \tag{3}$$

Let $\hat{Q}$ be the set of n-grams of a passage p composed using the terms of the question q. Therefore, we define $\hat{P} = \{x_1, x_2, ..., x_M\}$ as a sorted subset of $\hat{Q}$ that fulfills the following conditions:

1. $\forall x_i \in \hat{P} : h(x_i) \geq h(x_{i+1}) \quad i \in \{1, 2, ..., M-1\}$
2. $\forall x, y \in \hat{P} : x \neq y \Rightarrow T(x) \bigcap T(y) = \emptyset$
3. $\min\limits_{x \in \hat{P}} h(x) \geq \max\limits_{y \in \hat{Q} - \hat{P}} h(y)$

where $T(x)$ is the set of terms of the n-gram x, and $h(x)$ is the function defined by Formula 1.

The simplest measure of distance between two n-grams can be defined as the number of terms between them. However, this function has the disadvantage that it grows linearly and, therefore, the weight of the n-gram decreases too fast with respect to its distance from the heaviest n-gram. In order to address this issue, we use a logarithmic distance instead of the linear one. The distance function we have used is the following:

$$d(x, x_{max}) = 1 + k \cdot \ln(1 + L) \tag{4}$$

where L is the number of terms between the n-gram x_{max} (x_{max} is the n-gram with the maximum weight calculated in the Formula (1)) and the n-gram x of the passage. We have introduced the k constant to adjust the importance of the distance in the similarity formula. In previous experiments (Gómez, 2007), we have determined that the best value for this value is 0.1. The other added constants are used to avoid obtaining the infinity value when L is equal to 0.

Figure 3 presents an example. The first passage contains one question n-gram, and its similarity value is simply the sum of its terms divided by the sum of the weights of all question terms. However, the second passage has two question n-grams. The greatest n-gram is "*the Croatia*" with a weight of 0.6. The other question n-gram is "*capital of*" with a weight of 0.3 and a distance to the greatest n-gram of 7. If we calculate the similarity value for both passages, we obtain a similarity value of 0.9 for the first passage and a similarity value of 0.7 for the second one.

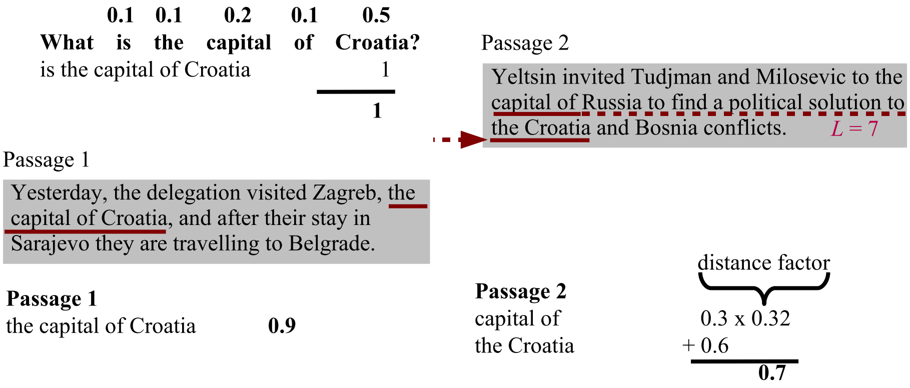
0.1 0.1 0.2 0.1 0.5
What is the capital of Croatia?
is the capital of Croatia 1
1
Passage 2
Yeltsin invited Tudjman and Milosevic to the capital of Russia to find a political solution to the Croatia and Bosnia conflicts.
L = 7
Passage 1
Yesterday, the delegation visited Zagreb, the capital of Croatia, and after their stay in Sarajevo they are travelling to Belgrade.
Passage 1
the capital of Croatia 0.9
Passage 2
capital of
the Croatia
distance factor
0.3 x 0.32
+ 0.6
0.7

Fig. 3 Example of the Distance Density n-gram model

In the *Distance Density n-gram* model, the term weights acquire significance with respect to the n-gram weights. Therefore, if an n-gram does not contain any of the relevant terms, this n-gram receives a much smaller weight than another one that includes such a term. Those n-grams that do not contain an irrelevant term (e.g., a stopword) will have weights that are only very slightly reduced. Another characteristic of this model is that the similarity value is not affected by the term permutations. That is, the n-gram "*is the capital of Croatia*" is given the same weight as the n-gram "*the capital of Croatia is*" because it is composed of question terms. This aspect is very important for languages whose expressions containing the answer are usually formulated by means of permutations of question terms.

3.2 Opinion Mining

Regarding the identification and classification of the subjective content of the documents, we proposed two Opinion Mining (OM) approaches, which are summarized in Figures 4 and 5.

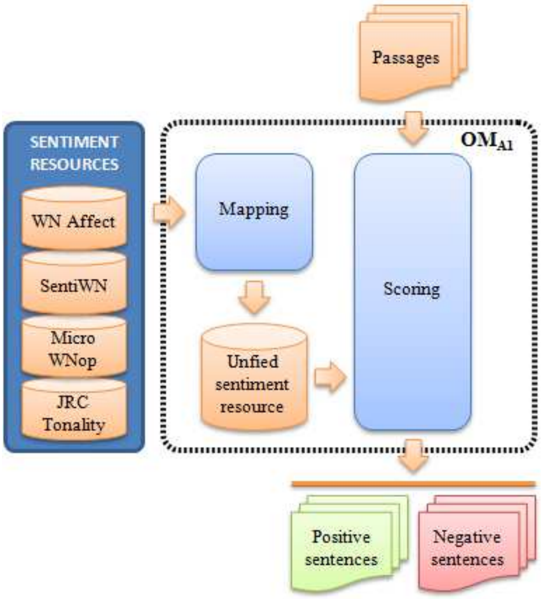

Fig. 4 First approach of the opinion mining module

In our first approach for OM (Figure 4), the initial step is to determine the opinionated sentences, assigning 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 opinions in a general context, in our initial approximation (**OM**$_{A1}$), we employed a simple, yet efficient method, presented in (Balahur et al, 2009d).

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) and the JRC tonality lists (Balahur et al, 2009d). 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). Such mapping was done differently depending on the resource dealt with. For instance, in WordNet Affect, the emotions anger and disgust were mapped onto the high negative class; the sadness emotion onto the negative; surprise onto the positive and finally joy was mapped onto the high positive class. As shown in (Balahur et al, 2009d), these values performed better than the usual assignment of only positive (1) and negative (-1) values. Additionally, Balahur et al. (2010d) show that this method can achieve up to 81% accuracy in classifying sentiment expressed in news. First, the score of each of the passages 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 final negative score leads to the system classifying the post as negative. Subsequently, we performed sentence splitting using Lingpipe[7] and classified the sentences we thus obtained according to their polarity, by adding the individual scores of the affective words (i.e., words that express sentiment) identified.

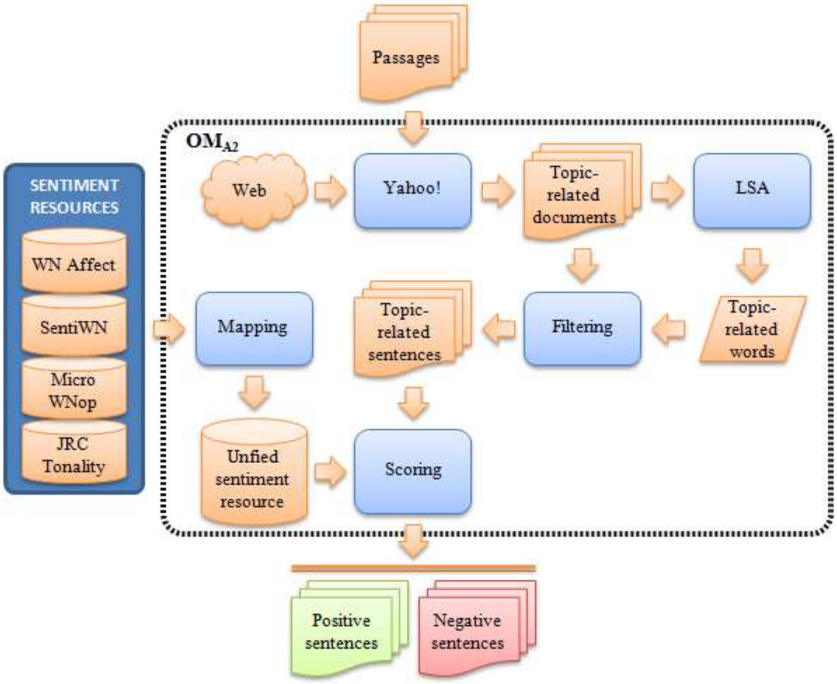

Fig. 5 Second approach of the opinion mining module

[7] http://alias-i.com/lingpipe/ (*Last Access: 06/02/2012*)

In the second approach ($\mathbf{OM}_{A2}$), the first step is to retrieve the 30 most related documents to a specific topic using the Web (in particular, we use Yahoo![8] search engine). Once the topic-related documents have been identified, Latent Semantic Analysis (LSA) through the Infomap NLP Software[9] is applied to identify the topic-related words. Subsequently, the sentences not containing such words will be filtered out, thus indicating that they are not associated with the topic discussed. Further on, the remaining sentences will be scored in the same manner as in the previous approach ($\mathbf{OM}_{A1}$), thus determining its polarity.

3.3 Text Summarization

As Text Summarization (TS) component, we integrated COMPENDIUM (Lloret et al, 2011) in the proposed framework. The selection of this summarization system was due to the fact that it was extensively evaluated with respect to different domains in previous research (Lloret, 2011), obtaining very competitive results compared with other state-of-the-art summarizers.

This TS system mainly relies on four stages for generating summaries: i) preprocessing; ii) redundancy detection; iii) relevance detection; and iv) summary generation. Firstly, a preprocessing is carried out in order to prepare the text for further processing. Once redundant information has been removed, a sentence is given a weight, indicating its relevance within the text. This weight will determine which sentences are to be selected and extracted. The effectiveness of such individual modules for summarization has been shown in previous research (Lloret et al., 2008), (Lloret and Palomar, 2009). Figure 6 depicts the summarization process, the stages of which are next explained in more detail.

3.3.1 Preprocessing

The preprocessing of the input document comprises sentence segmentation, tokenization, part-of-speech tagging, and *stopword* removal, and in each case external state-of-the-art tools and resources are employed. First of all, the text is segmented into sentences, which are the textual units considered for generating the summary. For this purpose, the sentence segmentation tool provided at DUC evaluation campaigns[10] is used. Further on, we identify each word of the text by means of a tokenizer[11] and we obtain its corresponding stem using the Porter Stemmer[12]. Then, a part-of-speech tagger (TreeTag-

[8] http://www.yahoo.com/ (*Last Access: 06/02/2012*)

[9] http://infomap-nlp.sourceforge.net/ (*Last Access: 06/02/2012*)

[10] http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz (*Last Access: 06/02/2012*)

[11] http://cogcomp.cs.illinois.edu/page/tools_view/8 (*Last Access: 06/02/2012*)

[12] http://tartarus.org/~martin/PorterStemmer/ (*Last Access: 06/02/2012*)

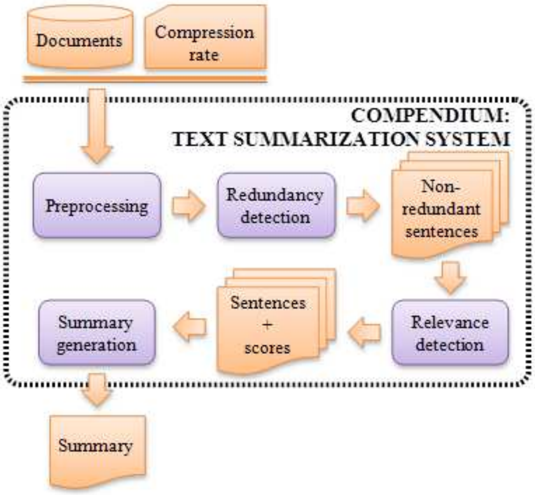

Fig. 6 COMPENDIUM text summarization process

ger[13]) assigns each word to its corresponding morphological category (noun, verb, adjective, preposition, adverb, determiner, pronoun, and conjunction). Finally, stop words are words which appear very frequent in documents, but do not carry any semantic information. This type of words are identified by using a specific list of English stop words[14].

3.3.2 Redundancy Detection

The aim of this stage is to avoid repeated information in the summary. Textual entailment is employed to meet this goal. A textual entailment relation holds between two text snippets when the meaning of one text snippet can be inferred from the other (Dagan et al, 2006). If such entailment relation can be identified automatically then it is possible to identify which sentences within a text can be inferred from others, as to avoid incorporating into summaries the sentences whose meaning is already included in the summary. In other words, the main idea here is to obtain a set of sentences from the text with no entailment relation, and then keep this set of sentences for further processing. In order to clarify this issue, the following example illustrate how the set of non-redundant sentences is obtained. For instance, let's assume that a document consists of a list of sentences:

$$S_1 \; S_2 \; S_3 \; S_4 \; S_5 \; S_6$$

[13] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (*Last Access: 06/02/2012*)

[14] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop (*Last Access: 06/02/2012*)

and we perform the entailment experiment as follows:

$NonRedundantSentences = \{S_1\}$
$NonRedundantSentences \longrightarrow entails \longrightarrow S_2 \Rightarrow NO$
$NonRedundantSentences = \{S_1, S_2\}$
$NonRedundantSentences \longrightarrow entails \longrightarrow S_3 \Rightarrow NO$
$NonRedundantSentences = \{S_1, S_2, S_3\}$
$NonRedundantSentences \longrightarrow entails \longrightarrow S_4 \Rightarrow YES$
$NonRedundantSentences = \{S_1, S_2, S_3\}$
$NonRedundantSentences \longrightarrow entails \longrightarrow S_5 \Rightarrow YES$
$NonRedundantSentences = \{S_1, S_2, S_3\}$
$NonRedundantSentences \longrightarrow entails \longrightarrow S_6 \Rightarrow NO$
$NonRedundantSentences = \{S_1, S_2, S_3, S_6\}$

Therefore, in this example, S_4 and S_5 are discarded from the text, and only the non-entailed sentences (i.e S_1, S_2, S_3 and S_6) are kept for further stages. To compute such entailment relations we have used the textual entailment approach presented in (Ferrández et al, 2007). This TE system relies on lexical (cosine similarity, Leveshtein distance), syntactic (dependency trees) and semantic measures based on WordNet Fellbaum (1998), and although its performance is around 60%, it has been shown in previous research (Lloret et al, 2008a), (Lloret et al, 2008b) that this technique is appropriate when addressing summarisation, for detecting redundant information.

3.3.3 Relevance Detection

The relevance detection module assigns a weight to each sentence, depending on how relevant it is within the text. This weight is based on the combination of two features: ***term frequency*** and the ***code quantity principle***. On the one hand, concerning term frequency, it is assumed that the more times a word appears in a document, the more relevant become the sentences that contain this word, following Luhn's idea (Luhn, 1958). On the other hand, the code quantity principle (Givón, 1990) is a linguistic theory which states that the less predictable information will be given more coding material. In other words, the most important information within a text will contain more lexical elements, and therefore it will be expressed by a high number of units (for instance, syllables, words or phrases). Noun-phrases within a document are flexible coding units that can vary in the number of elements depending on the level of detail desired for the information. Therefore, it is assumed that sentences containing longer noun-phrases are more important. The way the relevance of a sentence is computed is shown in Formula 5.

$$r_{s_i} = \frac{1}{\#NP_i} \sum_{w \in NP} |tf_w| \tag{5}$$

where:

$\#NPi$ = number of noun-phrases contained in sentence i,
tf_w = frequency of word w that belongs to a noun-phrase.

3.3.4 Summary Generation

Once the relevance score for each sentence is computed, the most important sentences (i.e. the ones with highest scores) up to a desired length are selected and extracted in the same order as they appeared in the original documents to form the final summary. The length of the summary is defined by a parameter given at the beginning of the TS process, which specifies either the compression rate for the summary with respect to the input, or the number of words we want the summary to have. Since the TS process employs Textual Entailment (TE) for detecting and removing redundancy, and term frequency (TF) together with the Code Quantity Principle (CQP) for computing the importance of each sentence in the relevance detection stage, we use the following nomenclature for referring to the resulting TS approach: TE+CQP+TF.

4 Experimental Framework

The objective of this section is to describe the corpus used and the experiments performed. Since the *Opinion Summarization Pilot* task in TAC 2008 provides a good environment to test and evaluate our approach, we are taking it as a basis, employing the same data and using the results of the participating systems for comparison purposes. This will allow us to analyze the benefits and limitations of our approach. Therefore, in this section, we first describe the data (Section 4.1). Further on, we focus on the explanation of the *Opinion Summarization Pilot* task guidelines (Section 4.2), and finally we outline the set of experiments we carried out (Section 4.3).

4.1 Corpora

For our experiments, we used the TAC 2008 data for the *Opinion Summarization Pilot* task[15]. Specifically, this data consisted of a subset of documents of the *Blog06* collection, which comprised 609 blogs clustered into related topics. In total, there were 25 different topics with 24 blogs related to them on average. The topics include people (e.g. George Clooney), events (e.g. Sheep and Wool Festival), companies (e.g. Starbucks coffee shops), products (e.g. Windows Vista), or TV shows (e.g. Mythbusters). Table 2 shows the complete list of topics we dealt with.

[15] http://www.nist.gov/tac/data/past/2008/OpSummQA08.html (*Last Access: 06/02/2012*)

Topic category	Topic
People	George Clooney
Companies	Carmax; Jiffy Lube; Starbucks coffee shops; Subway Sandwiches; Trader Joe's; Zillow
Computer Applications	Windows Vista; Picasa; YouTube
Organizations	UN Commission on Human Rights; NAFTA;World Bank
TV programmes	MythBusters; talk show hosts; women in Numb3rs
Festivals	Sheep and Wool Festival
Music	System of a Down
Books	A Million Little Pieces
Events (controversial topics)	architecture of Frank Gehry; tax breaks for hybrid automobiles; Whole Foods wind energy; China one-child per family law; David Irving's arrest in Austria for Holocaust denial criminalizing flag burning

Table 2 Topic categories of the blog collection.

4.2 Opinion Summarization Pilot task within TAC 2008

The goal of the *Opinion Summarization Pilot* task was to generate short coherent summaries of text. These summaries should contain the answers to opinion questions retrieved from blogs. In particular, given a set of 25 topics, a set of blogs from the Blog06 collection and a list of questions from the question answering track[16], participating systems had to produce a summary that answered these questions. The questions generally required determining opinion expressed on each individual topic. Additionally, a set of text snippets were also provided, which already contained the answers to the questions. These snippets were provided by real question answering systems, 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.

An example of three topics together with their corresponding questions is given in Table 3.

Regarding the length of the summaries, the TAC organization established a maximum length for the summaries, which could not exceed 7000 non-whitespace characters per question.

In this research, we follow the TAC guidelines using our proposed unified framework with IR, OM and TS components, except those which related to the length of the summaries. Instead, we generated summaries of different compression rates in order to analyze whether we could find one that is particularly suitable for this task. Furthermore, as input for our approach, we only used the blog data collection, the topics and the questions which we have to find the answers from. As opposed to the participants in TAC 2008, we

[16] http://www.nist.gov/tac/data/past/2008/OpSummQA08.html (*Last Access: 06/02/2012*)

Topic	Questions
Starbucks coffee shops	Why do people like Starbucks better than Dunkin Donuts? Why do people like Dunkin Donuts better than Starbucks?
Windows Vista	What features do people like about Vista? What features do people dislike about Vista?
George Clooney	Why do people like George Clooney? Why do people dislike George Clooney?

Table 3 Example of TAC 2008 topics and questions.

did not use the snippet list containing the answers to the questions. Finally, in order to provide a deeper analysis of how our approach performs, we experimented with different coneations for the information retrieval and opinion mining components, which are next explained.

4.3 Experiments

The proposed framework is highly modular. This leads to a very suitable framework, where different IR, OM, and TS systems could be analyzed on their own, as well as in combination with the others. Furthermore, each individual component of these systems can be tuned by choosing among different parameters and options. Therefore, we first describe the parameters for each of the components[17], and further on we explain all the combinations we have analyzed.

The specific configurations of each of the aforementioned components are:

- **Information Retrieval:** For the scope of this paper, we analyze two different passage lengths: 1 and 3. This refers to the number of sentences taken from each document retrieved by the IR system (1 means that we only output a sentence per retrieved document and query, whereas by considering a 3-length snippet the IR component will retrieve 3 sentences per document and per query). In previous research works ((Gómez, 2007), (Balahur et al, 2010b), (Balahur et al, 2010a)), these passage lengths have been proven to be the most appropriate.
- **Opinion Mining:** As far as the opinion mining component is concerned, we analyze two approaches: OM_{A1} and OM_{A2}, as explained in Section 3.2.
- **Text Summarization:** COMPENDIUM will generate summaries using the TE+CQP+TF approach. This means that redundant information is removed first by means of a textual entailment module, and then the most relevant sentence are determined by relying on the frequency of words and the code quantity principle following the approach explained in Section 3.3.

[17] For specific detail of the different IR, OM and TS components, please refer to Section 3.

Regarding to the length of the summaries, different compression rates summaries (from 10% to 50%) are produced, instead of having a fixed length as in the TAC competition.

Moreover, the approaches analyzed comprise:

- **IR-TS:** This combination only uses the information retrieval and the text summarization components. Once the information retrieval sub-system has obtained the most relevant passages, the summarization sub-system takes them as input and determines which information is the most relevant, thus obtaining the final summary. Here, no OM is employed, and taking into account that we experimented with two different passage length, we end up with two different approaches: $\boldsymbol{IR_{p1}\text{-}TS}$ and $\boldsymbol{IR_{p3}\text{-}TS}$.
- **IR-OM:** This approach differs from the previous one in that it uses OM and not TS. First, the most relevant passages are retrieved by the IR module, as in the aforementioned approach, and then the subjective information is found and classified within them using the OM approach previously described in Section 3.2. Finally, the summary is generated by concatenating the top n sentences of subjective information. The n is computed according to the compression rate desired (e.g., if the source document had 100 sentences, and we are producing a 50% compression ratio summary, we would produce the summary by extracting the 50 highest relevant sentences). As before, we have two passage lengths, and in this case, we also have two OM approaches, leading to: $\boldsymbol{IR_{p1}\text{-}OM_{A1}}$; $\boldsymbol{IR_{p1}\text{-}OM_{A2}}$; $\boldsymbol{IR_{p3}\text{-}OM_{A1}}$; and $\boldsymbol{IR_{p3}\text{-}OM_{A2}}$.
- **IR-OM-TS:** Finally, in this approach, the unified framework is tested. The process is the same as the IR-OM approach, but then at the end we integrate the TS component, to select and extract the most relevant opinionated facts from the pool of subjective information identified by the OM component. Four different approaches result from the integration of the three components: $\boldsymbol{IR_{p1}\text{-}OM_{A1}\text{-}TS}$; $\boldsymbol{IR_{p1}\text{-}OM_{A2}\text{-}TS}$; $\boldsymbol{IR_{p3}\text{-}OM_{A1}\text{-}TS}$; and $\boldsymbol{IR_{p3}\text{-}OM_{A2}\text{-}TS}$.

Moreover, apart from these approaches, two baselines were also defined. On the one hand, a baseline using the list of snippets provided by the TAC organization was also established. This baseline produces a summary by joining all the answers in the snippets that related to the same topic, and it will be referred as a ***QA-snippets***. On the other hand, we took as a second baseline the approach from our participation in TAC 2008. Such baseline did not take into account any information retrieval or question answering system to retrieve the fragments of information which may be relevant to the query. In contrast, this was performed by computing the cosine similarity between each sentence in the blog and the query. The similarity score was computed with Pedersen's Text Similarity package[18]. After all the potential relevant sentences for the query were identified, they were classified in terms of subjectivity and polarity, and the ones with highest values of polarity intensity were selected for the final

[18] http://www.d.umn.edu/~tpederse/text-similarity.html (*Last Access: 06/02/2012*)

Approach	**Description**
QA-snippets	baseline; no IR; no OM; no TS
DLSIUAES	baseline; TAC participation
IR$_{p1}$-TS	IR passage length=1; no OM; and TS=TE+CQP+TF
IR$_{p3}$-TS	IR passage length=3; no OM; and TS=TE+CQP+TF
IR$_{p1}$-OM$_{A1}$	IR passage length=1; OM=lexica; and no TS
IR$_{p1}$-OM$_{A2}$	IR passage length=1; OM=topics; and no TS
IR$_{p3}$-OM$_{A1}$	IR passage length=3; OM=lexica; and no TS
IR$_{p3}$-OM$_{A2}$	IR passage length=3; OM=topics; and no TS
IR$_{p1}$-OM$_{A1}$-TS	IR passage length=1; OM=lexica; and TS=TE+CQP+TF
IR$_{p1}$-OM$_{A2}$-TS	IR passage length=1; OM=topics; and TS=TE+CQP+TF
IR$_{p3}$-OM$_{A1}$-TS	IR passage length=3; OM=lexica; and TS=TE+CQP+TF
IR$_{p3}$-OM$_{A2}$-TS	IR passage length=3; OM=topics; and TS=TE+CQP+TF

Table 4 Description of the approaches tested

summary. We name this baseline as ***DLSIUAES***. Table 4 summarizes briefly all the different approaches tested.

In the next section, the results obtained together with a detailed analysis and discussion is provided.

5 Evaluation Methodology

The goal of this section is to show and analyze the results obtained from the experimentation. Therefore, the benefits and limitations of the proposed Opinion Retrieval, Mining and Summarization framework can be analyzed.

Although we used the same corpus as in the *Opinion Summarization Pilot* task, and we followed similar guidelines, the evaluation we propose differs slightly from the one carried out in the competition. The reason for opting for another evaluation is that the evaluation carried out in TAC had some limitations, and therefore was not suitable for our purposes. Such limitations are analyzed in detailed in Section 5.1.1. In this manner, although our evaluation is mainly based on the Gold Standard nuggets provided within the TAC 2008 *Opinion Summarization Pilot* task, we have also created a new version of the Gold Standard, by adding new snippets of text containing relevant answers to the proposed opinion questions.

In this section, all the issues concerning the evaluation are explained. These comprise the original evaluation method used in the *Opinion Summarization Pilot* task at TAC (Section 5.1), including the analysis of its drawbacks (Section 5.1.1), as well as the extended version for the evaluation method we propose (Section 5.1.2). Further on, the results obtained for the different configurations of our suggested framework are provided in Section 5.2, together with its comparison with the baselines and the results obtained by the TAC participants.

5.1 Nugget-based Evaluation at TAC

Within the *Opinion Summarization Pilot* task, each summary was evaluated according to its content using the Pyramid method (Nenkova et al, 2007). The goal of this method is to identify relevant information with the same meaning according to different human experts and determine what pieces of information on which they agree should be included in the summary. Each of these pieces is called a Summary Content Unit (SCU) or nugget, and it has a weight assigned depending on the number of human assessors who agreed that they are important to the summary. This weight indicates the relevance of the nugget for the summary (the higher weight, the more important the information is).

Within the scope of the *Opinion Summarization Pilot* task, a group of human assessors who were experts in summarization manually built a list of nuggets. Then, the assessors who were in charge of evaluating the summaries in the TAC conference, used such list to count how many of them were present in the automatic summaries and summed up their weights. The final step was to obtain the values for recall, precision and F-measure, defined as follows:

$$Recall = \frac{total\ weights\ of\ matched\ nuggets}{total\ weights\ of\ all\ nuggets}$$

$$Precision = \frac{number\ of\ matches * 100}{length\ of\ the\ summary}$$

$$F - measure_{\beta} = 1 = \frac{2 * Recall * Precision}{Recall + Precision}$$

Several examples of nuggets corresponding to different topics can be seen in Table 5; the weight of each nugget is shown between brackets.

Topic	Nugget (weight)
Carmax	CARMAX prices are firm, the price is the price (0.9)
Jiffy Lube	They should have torque wrenches (0.2)
Talk show hosts	Funny (0.78)

Table 5 Example of evaluation nuggets and weights for *Carmax*, *Jiffy Lube*, and *talk show hosts* topics, respectively.

For our evaluation, we took as a starting point the list of nuggets provided in the TAC conference, in order to distinguish between the essential and non-essential information that the summary should capture. However, after analyzing in more detail the list of nuggets, we found some limitations, which are explained in the next section.

5.1.1 Limitations of the Nugget Evaluation

The evaluation method suggested at TAC requires a lot of human effort to identify the relevant fragments of information (nuggets) and compute how many of them a summary contains, resulting in a very costly and time-consuming task. This is a general problem associated with the evaluation of summaries, which makes the task of summarization evaluation especially difficult. Nevertheless, even if we were to disregard the inherent difficulties mentioned, we also detected additional problems, which will be explained in the following sections.

The average number of nuggets for each topic is 27, and the number of total average characters considering all the nuggets related to a topic is 1931. Given this fact, the 7000 characters limit of the summaries, as well as the evaluation methodology, by which irrelevant info is penalized, we can easily see that longer summaries are penalized, even if they contain all the relevant nuggets. After analyzing in detail all the provided nuggets, we mainly classified the possible problems into six groups, which are:

1. **Some of the nuggets were expressed differently from how they appeared in the original blogs.** Since most of the summarization systems are extractive, this fact led to the fact that only a manual evaluation was possible; otherwise it would have been very difficult to account for the presence of such nugget in the summary (since they are not using the same vocabulary as the orginal blogs). To make the evaluation more difficult, there were some cases, where these nuggets were assigned a higher score, as for instance *"Buying from CARMAX is low stress"* which had a weight of 0.7 .
2. **Some nuggets for the same topic express the same idea, despite their not being identical.** In these cases, we are counting a single piece of information in the summary twice. This happens for example to the nuggets *"NAFTA needs to be renegotiated to protect Canadian sovereignty"* and *"Green Party: Renegotiate NAFTA to protect Canadian Sovereignty"*, belonging to the topic *Nafta.*
3. **The meaning of one nugget can be deduced from another's**, which is also related to the problem stated before. For the topic *"Subway sandwiches"*, two of the nuggets are *"reasonably healthy food"* and *"sandwiches are healthy"*).
4. **Some of the nuggets do not appear in context and thus their meaning is not completely clear/unambiguous** (e.g. *"hot"*, *"fun"*). This fact implies that a summary might include such terms in a different context, and the evaluation would incorrectly positively reward the summary, when it is not.
5. **A sentence in the original blog can be covered by several nuggets**. For instance, both nuggets *"it is an honest book"* and *"it is a great book"* correspond to the same sentence *"It was such a great book- honest and hard to read (content not language difficulty)"*. In this case, it is not clear how

to proceed with the evaluation; whether to count both nuggets or only one of them.

6. **Additional information which is also relevant for the topic is not present in any nugget contained in the Gold Standard**. We can find an example of this in the following sentence: *"I go to Starbucks because they generally provide me better service"*. Although it is relevant with respect to the topic and it appears in a number of summaries, it would be not counted because it has not been included by the TAC human annotators on the list of relevant nuggets.

5.1.2 Extended Nugget-based Evaluation

Having observed the drawbacks of the evaluation methodology proposed in the TAC 2008 Opinion Pilot task and given our stated objective of evaluating an extensive number of approaches, we propose a new evaluation methodology. This new assessment method is based on the extension of the initial list of nuggets proposed in the evaluation methodology.

The underlying idea is to create an extended set of nuggets that serve as a reference for assessing the content of the summaries. In this manner, we can map each original nugget with the set of sentences in the original blogs that are most similar to it, thus generating an extended gold-standard summary for each topic. In order to create this extended gold-standard list of nuggets, we compute the cosine similarity[19] between every nugget and all the sentences in the blog related to the same topic.

After analysing possible similarity thresholds (from 0 to 1), we established a threshold of 0.5, meaning that if a sentence is equal or above such similarity value, it is also relevant. The reason for such threshold was that when using higher values we hardly obtained similar sentences, since the tool we were employing only relied on lexical similarity. In a similar manner, when selecting thresholds below 0.5, the number of similar sentences increased, but after revising them manually we realized that we were incorrectly considering sentences as similar. Therefore, 0.5 was the threshold that allow the identification of potential similar sentences. However, one of the main disadvantages of this threshold value is that we may end up considering as relevant the sentences that share the same vocabulary, but in fact are not relevant to the summary. In order to avoid this pitfall, once we had identified the entire set of sentences above the 0.5 threshold of similarity to each nugget, we carried out a manual analysis and discarded the non-similar ones.

Having created the extended set of nuggets, we grouped the ones pertaining to the same topic, which were considered as a gold-standard summary. After having performed this semi-automatic process of relevant nugget detection, the average number of nuggets per topic is 53, doubling the number of original nuggets provided at TAC 2008. Subsequently, the main objective of the new

[19] The cosine similarity was computed using Pedersen's Text Similarity Package: http://www.d.umn.edu/~tpederse/text-similarity.html (*Last Access: 06/02/2012*)

evaluation methodology is to overcome the shortcomings identified, as far as information content is concerned, and the manual effort required for the final assessment.

Further on, our summaries are compared against this new gold-standard using ROUGE (Lin, 2004). ROUGE is a state-of-the-art tool for evaluating summaries automatically. Basically, this tool computes the number of different kinds of overlap n-grams between an automatic summary and a human-made summary. For our evaluation, we compute ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE-SU4 (it measures the overlap of skip-bigrams between a candidate summary and a set of reference summaries with a maximum skip distance of 4), and ROUGE-L (Longest Common Subsequence between two texts). The results are presented and discussed in the following section.

5.2 Results and Discussion

This section contains the results obtained for the approaches tested within our Opinion Retrieval, Mining and Summarization framework. Firstly, we show and analyze the results of our different approaches, and consequently we compare the best-performing one with the baselines, the average results obtained by the participants in the *Opinion Summarization Pilot* task, as well as the best and worst performing TAC systems. In this manner, we will be able to account for the strengthens and weaknesses of our proposed approach compared to the state of the art.

Table 6 presents the results obtained by all our approaches in terms of precision (Pre), recall (Rec) and F-measure ($F_\beta = 1$) at the different summarization compression rates we experimented with. These approaches comprise in the first place the combination of two of the framework components at a time, and then the whole framework, with its three components (IR, OM and TS).

Generally speaking, the results obtained show better figures for precision than for recall, and therefore the F-measure value, which combines both values, will be also influence by these values. In some cases, for instance in the approach IR_{p3}-TS for a 20% compression rate, the value of recall is much lower than the precision, thus affecting significantly the final F-measure. In constrast, the opposite rarely happens, although we can also find some cases where the recall is better than precision (e.g.IR_{p3}-OM_{A1}-TS for 40% and 50% compression rates). This can be explained by the low compression rate, which means that the probability of relevant data to be obtained in the summary is higher. Good precision values means that the information selected by our approaches is the correct one, despite it not including all the relevant data.

As it was expected, our best performing approach is the one which integrates all the components (i.e., information retrieval, opinion mining and text summarization) and a passage length of 3 is taken into consideration. This is in line with the research carried out in Balahur et al. ((2009a), (2009b),(2010a), (2010b)) that have shown that 3-sentence-long snippet retrieval for opinion

Approach		Summary length (compression rate)				
Name	**ROUGE-1 (%)**	**10%**	**20%**	**30%**	**40%**	**50%**
IR$_{p1}$-TS	Pre	21.14	24.56	30.82	33.35	**35.47**
	Rec	14.20	17.46	21.12	24.42	26.52
	F$_{\beta=1}$	16.20	19.27	22.94	25.67	27.41
IR$_{p3}$-TS	Pre	23.17	30.17	**32.80**	**33.80**	33.00
	Rec	17.39	19.95	22.98	25.28	27.01
	F$_{\beta=1}$	18.94	21.42	23.31	24.01	23.87
IR$_{p1}$-OM$_{A1}$	Pre	16.30	14.87	25.32	28.96	31.96
	Rec	8.16	20.97	18.24	24.81	27.68
	F$_{\beta=1}$	10.48	16.78	20.61	26.34	29.08
IR$_{p1}$-OM$_{A2}$	Pre	11.55	13.62	18.09	22.34	25.00
	Rec	5.46	10.97	13.73	18.78	21.16
	F$_{\beta=1}$	7.30	11.97	15.28	19.92	22.45
IR$_{p3}$-OM$_{A1}$	Pre	19.69	26.09	26.43	32.70	24.75
	Rec	10.84	21.60	**28.86**	25.93	35.02
	F$_{\beta=1}$	13.49	23.03	27.12	28.00	27.91
IR$_{p3}$-OM$_{A2}$	Pre	11.58	13.70	21.95	25.97	23.46
	Rec	9.46	12.30	21.91	24.56	29.58
	F$_{\beta=1}$	12.67	12.70	21.48	23.97	25.16
IR$_{p1}$-OM$_{A1}$-TS	Pre	24.29	26.17	29.73	30.82	32.54
	Rec	14.45	18.58	22.32	23.63	26.32
	F$_{\beta=1}$	16.53	20.65	24.58	25.75	28.12
IR$_{p1}$-OM$_{A2}$-TS	Pre	24.29	26.17	29.73	30.82	32.54
	Rec	16.90	20.02	23.36	24.15	26.77
	F$_{\beta=1}$	19.45	22.13	25.36	25.94	28.40
IR$_{p3}$-OM$_{A1}$-TS	Pre	27.27	30.18	30.91	30.05	30.19
	Rec	20.56	**24.76**	28.25	**31.67**	**34.47**
	F$_{\beta=1}$	22.65	**26.23**	**27.98**	**29.18**	**29.74**
IR$_{p3}$-OM$_{A2}$-TS	Pre	**30.16**	**32.11**	32.35	32.41	32.11
	Rec	**20.64**	24.03	27.25	29.78	32.68
	F$_{\beta=1}$	**23.28**	25.64	27.42	28.44	29.21

Table 6 Results of our IR-OM-TS approaches

question answering[20] performs better than the traditional one-sentence-long snippet retrieval. This approach is *IR$_{p3}$-OM$_{A1}$-TS*, and as far as OM is concerned, the approach dealing with sentiment lexica seems more suitable than the one which uses LSA.

Analyzing the individual results, we subsequently try to determine the reasons why our approach performs better using some approaches and not so good when employing others. Concerning the IR component, it is important to mention that a passage length of 1 always obtains poorer results than when it is increased to 3, meaning that the longer the passage, the better. AS we have already mentioned, this conclusion is also supported in (Balahur et al, 2009b), (Balahur et al, 2010b), (Balahur et al, 2010a). When IR is used in conjunction with the TS approach, the results increase for higher compression rates (10%-30%) for the cases where a passage length of 1 is used, in comparison to combining only IR and OM. However, for lower compression rates (40%-50%) as well as when a longer passage length is selected (i.e., 3),

[20] We can consider opinion question answering as a specific type of information retrieval.

using OM influences positively in the overall performance. On the one hand, when not taking into account an OM component, TS is not capable of distinguishing whether the information expresses a positive or a negative sentiment, selecting only the most important sentences according to the techniques employed, thus disregarding the relevance of the snippet as far as the required answer polarity is concerned. On the other hand, if TS is not employed, and we simply rely on the sentences with highest opinion intensity values to build up the final summary, it may happen that the sentences chosen are not the most important ones, thus leading to poor results in some cases compared to the same approach but using TS instead of OM (e.g. IR_{p3}-OM_{A1} compared to IR_{p3}-TS).

Furthermore, in order to account for the significance of the results, we performed a t-test at a 95% confidence level. We compared our best approaches (i.e., IR_{p3}-OM_{A1}-TS and IR_{p3}-OM_{A2}-TS) with the remaining ones. Our findings show that out that in the case of the IR_{p3}-OM_{A1}-TS approach the results were statistically significant for all the cases at most of the compression rates (except the 10% one). At the 10% compression rate, the IR_{p3}-OM_{A2}-TS approach performs significantly better than any of the other proposed approaches.

Regarding the best summary length, we observed that in general terms, the more content we allow for the summary, the better. In other words, compression rates of 50% get higher results than 20% or 10%. However, it is worth mentioning that this does not occur when we combine IR using a passage length of 3 sentences with TS, without using any OM component in between. In these cases, 40% is the optimum compression ratio for the precision and F-measure values. As it can be seen, compression rates of 50% for these values decrease a little bit with respect to the ones for 40%. In this case, the fact of introducing additional sentences may be result in an increase of noisy information in the summary.

Finally, it is worth stressing upon the fact that we are facing a very challenging task. This is partly due to the nature of the source documents (i.e., blogs), which contain a lot of irrelevant and noisy information, such as advertisements, or comments completely unrelated to the topic that is being discussed. Although the results in themselves are not very high (around 30%), they are in line with the state of the art, as can be seen in Table 7, where our best performing approach, from all compression rates analyzed, is compared with respect to other approaches. In addition to the two baselines (QA-Snippets and DLSIUAES), we also compute the performance of the TAC participants, and we show the results of best and worst systems, as well as the average results of the competitions. It is worth mentioning that, apart from considering all the TAC participants together, we have also distinguished those systems that did not use the initial set of snippets provided by the TAC organization, and we have denoted them by TAC'. The reason was because such systems did not have the ideal information from which generate the summaries, and therefore they follow an approach that is more similar to ours.

Approach		**Performance (ROUGE)**			
Name	**%**	**R-1**	**R-2**	**R-L**	**R-SU4**
IR$_{p3}$-OM$_{A1}$-TS (50%)	Pre	30.19	7.34	29.00	11.37
	Rec	34.47	8.31	33.24	12.76
	F$_{\beta=1}$	**29.74**	**7.22**	**28.60**	**11.13**
QA-snippets	Pre	17.97	8.76	17.65	9.98
	Rec	71.24	31.30	70.10	37.44
	F$_{\beta=1}$	24.73	11.58	24.29	13.45
DLSIUAES	Pre	20.54	7.00	19.46	9.29
	Rec	57.66	18.98	54.61	25.77
	F$_{\beta=1}$	27.04	9.10	25.59	12.22
Best TAC (system 13)	Pre	40.16	20.81	37.65	22.35
	Rec	56.65	27.01	53.66	39.68
	F$_{\beta=1}$	**41.39**	**20.72**	**38.87**	**22.36**
Average TAC	Pre	23.74	8.35	22.72	10.81
	Rec	56.65	19.37	54.56	25.40
	F$_{\beta=1}$	27.45	9.64	26.33	12.46
Best TAC' (system 16)	Pre	28.68	9.81	27.98	13.14
	Rec	53.34	16.12	51.96	22.57
	F$_{\beta=1}$	**34.23**	**11.24**	**33.35**	**15.24**
Average TAC'	Pre	20.42	6.06	19.55	8.62
	Rec	56.45	17.3	54.40	24.11
	F$_{\beta=1}$	**24.31**	**7.25**	**23.31**	**10.29**
Worst TAC and TAC' (system 10)	Pre	11.37	3.55	11.10	5.08
	Rec	72.17	20.95	70.89	30.93
	F$_{\beta=1}$	18.33	5.65	17.92	8.13

Table 7 Comparison with other systems

As it was previously stated, our best performing approach is the one which employs all the components in the unified framework (i.e., information retrieval, opinion mining and summarization) using a passage length of 3, and sentiment lexica for identifying and classifying the opinion found without filtering according to topic relevance (IR_{p3}-OM_{A1}-TS). Although the compression rate which obtains best results is not very high (50%), the final summaries have an average length of 2,333 non-white space characters. This is significantly low compared to the length established at the *Opinion Summarization Pilot* task, which was 7,000 non-white space characters per question. In TAC 2008 competition, the length of the final summary could reach up to 14,000 characters, because most of the times each topic had two questions that had to be answered in the same summary. Moreover, this is directly related to the results obtained. Whereas the results of TAC participants are much better for the recall value than ours, if we take a look at the precision, our approach obtains better results according to this value for some of the TAC results. The reason why the recall value is so high in the TAC participant systems is due to the length of the summaries. The longer a summary is, the more chances it has to contain information related to the topic. However, not all this information may be relevant, as it is shown in the results for the precision values, which decrease considerably compared to the recall ones.

Regarding the comparison between our approach and the two proposed baselines (*QA-snippets* and *DLSIUAES*), our approach performs significantly better at a 95% confidence level[21] than both baselines for F-measure in ROUGE-1 and ROUGE-L. However, as it can be seen, it obtains lower results for ROUGE-2 and ROUGE-SU4. This happens because the value for the recall in the *QA-snippets* and *DLSIUAES* baselines is much higher, and therefore, the F-measure will benefit from this situation. What ROUGE-2 and ROUGE-SU4 provides is the number of common bigrams between a model and an automatic summary. In this case, the summaries generated by the baselines contain a higher number of bigrams in common to the model nuggets than our approach. However, it is worth noting that our approach performs better when comparing the longest common subsequence (ROUGE-L).

The last five rows of the table show the average results for the TAC participant systems in the *Opinion Summarization Pilot* task. As we previously mentioned, the average, the best and worst TAC results computed correspond to two different scenarios. On the one hand, for the *Best TAC*, *Average TAC*, and *Worst TAC*, the summaries for all the participants have been re-evaluated using ROUGE in the same way as we did for our approaches, and then the average is obtained. At this point it is important to mention that the TAC organization provided a list of snippets[22] that already contained the answer to the questions for each topic. However, such list was optional, and therefore not all the systems used it. Consequently, we also decided to select those participant systems that did not use the optional snippets and evaluated the generated summaries. The average results are shown in the *Best TAC'*, *Average TAC'*, and *Worst TAC'* rows.

If we take the average results, it can be seen that, when the optional snippets are not used, the results decrease by approximately 11%, 25%, 11% and 17%, with respect to using them (*Average TAC* vs. *Average TAC'*), according to the F-measure for ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4, respectively.

On the one hand, comparing our approach (IR_{p3}-OM_{A1}-TS) with respect to the TAC' participants, we would like to stress that our precision outperforms the one of the Best TAC' system (system 16), except for ROUGE-2 and ROUGE-SU4 metrics. However, although we could not improve the F-measure with respect to this approach, we improved it with respect to the results obtained by the average TAC' participants by 13.20%, and this improvement was statistically significant, for all the individual ROUGE metrics, except for ROUGE-2, where the results are almost identical.

On the other hand, considering all the TAC systems together, although our approach is not capable to perform as good as the best TAC system (system 13), our F-measure performance is higher than the one obtained averaging all TAC participants (except for ROUGE-2 and ROUGE-SU4).

[21] A t-test was carried out in order to account the significance of the results.

[22] We have used these snippets for building our *QA-snippets* baseline.

Both, the best performing system in each category, system 13 and system 16, were very competitive, thus improving by 11% and 12%, on average, the results of the second best systems of the *Opinion Summarization Pilot* task, respectively.

In general, our approach was above the average, surpassing the average results of the TAC conference, although not reaching as very good results than the best performing systems in both cases. Since there were 36 participating systems, our approach would have obtained the following rankings in the TAC competition: 17/36 (ROUGE-1); 23/36 (ROUGE-2); 15/36 (ROUGE-L); and 19/36 (ROUGE-SU4). When considering only the systems that did not use the snippets provided within the conference, we would have achieved much better positions: 4/19 (ROUGE-1); 9/19 (ROUGE-2); 3/191 (ROUGE-L); and 8/19 (ROUGE-SU4). Therefore, our unified framework performs according to the state of the art, and it can be considered competitive enough with respect to other systems.

6 Conclusions and Future Work

This article presented a unified framework for Opinion Retrieval, Mining and Summarization. Our proposed framework integrates three different components (i.e., information retrieval, opinion mining and text summarization) in order to generate opinion summaries from subjective texts found on the Web, in particular blogs. These components are crucial in the task of opinion summarization, since our final goal is to provide users with the correct information.

Our analysis comprises different configurations and approaches: i) varying the length for retrieving the passages of the documents in the retrieval information stage; ii) studying a method that takes into consideration sentiment lexica for detecting and classifying opinions in the retrieved passages and comparing it to another that uses topic-sentiment identification by means of LSA; and iii) generating summaries of different compression rates (10% to 50%). The results obtained showed that the proposed methods are appropriate to build the Opinion Retrieval, Mining and Summarization framework, being in line with the state-of-the-art approaches, and surpassing the average TAC' participants by 13.20% approximately (F-measure for ROUGE-1). From the evaluation performed and the results obtained, we can conclude that our approach is competitive enough to be used for generating opinion-oriented summaries within a single process that integrates information retrieval as well as opinion mining.

However, we also could notice that there may be better approaches, and therefore, in the future, we plan to continue improving the individual components for the Opinion Retrieval, Mining and Summarization framework, as well as investigating other suitable approaches that can enrich the proposed framework. Moreover, it would be very interesting to analyze the performance of the framework, not only in blogs, but also with other types of texts of different nature, such as books, reviews or newspapers. In the long term our final aim to exploit and apply the framework in real contexts, so both individuals and organizations can benefit from these technologies.

Acknowledgements

This research work has been funded by the Spanish Government through the project TEXT-MESS 2.0 (TIN2009-13391-C04) and by the Valencian Government through projects PROMETEO (PROMETEO/2009/199) and ACOMP/2011/001.

References

Aslandogan YA, Yu CT (1999) Techniques and Systems for Image and Video Retrieval. IEEE Transactions on Knowledge and Data Engineering 11(1):56–63, DOI dx.doi.org/10.1109/69.755615

Balahur A, Lloret E, Ferrández O, Montoyo A, Palomar M, Muñoz R (2008) The DLSIUAES Team's Participation in the TAC 2008 Tracks. In: Proceedings of the Text Analysis Conference (TAC)

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2009a) A Comparative Study of Open Domain and Opinion Question Answering Systems for Factual and Opinionated Queries. In: Proceedings of the International Conference RANLP-2009, Association for Computational Linguistics, Borovets, Bulgaria, pp 18–22

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2009b) Opinion and Generic Question Answering Systems: A Performance Analysis. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 157–160

Balahur A, Kabadjov M, Steinberger J, Steinberger R, Montoyo A (2009c) Summarizing Opinions in Blog Threads. In: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC), City University of Hong Kong Press, pp 606–613

Balahur A, Steinberger R, van der Goot E, Pouliquen B, Kabadjov M (2009d) Opinion Mining from Newspaper Quotations. In: Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content at the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp 523–526

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2010a) Going beyond Traditional QA Systems: Challenges and Keys in Opinion Question Answering. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 27–35

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2010b) Opinion Question Answering: Towards a Unified Approach. In: Proceeding of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence, IOS Press, Amsterdam, The Netherlands, The Netherlands, pp 511–516

Balahur A, Kabadjov M, Steinberger J (2010c) Exploiting Higher-level Semantic Information for the Opinion-oriented Summarization of Blogs. International Journal of Computational Linguistics and Applications 1(1-2):45–59

Balahur A, Steinberger R, Kabadjov MA, Zavarella V, der Goot EV, Halkia M, Pouliquen B, Belyaeva J (2010d) Sentiment Analysis in the News. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp 2216–2220

Beineke P, Hastie T, Manning C, Vaithyanathan S (2003) An Exploration of Sentiment Summarization. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, Stanford, US, pp 1–4

Bossard A, Généreux M, Poibeau T (2008) Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions. In: Proceedings of the Text Analysis Conference (TAC)

Brin S, Page L (1998) The Anatomy of a Large-scale Hypertextual Web Search Engine. In: Proceedings of the seventh international conference on World Wide Web 7, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, WWW7, pp 107–117

Buscaldi D, Rosso P, Gómez-Soriano JM, Sanchis E (2010) Answering questions with an n-gram based passage retrieval engine. Journal of Intelligent Information Systems 34:113–134

Carenini G, Cheung JCK (2008) Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality. In: Proceedings of the Fifth International Natural Language Generation Conference, ACL 2008, Ohio, USA, pp 33–40

Cerini S, Compagnoni V, Demontis A, Formentelli M, Gandini G (2007) Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In: Sansó A (ed) Language resources and linguistic theory: Typology, second language acquisition, English linguistics, Franco Angeli, Milano, IT, pp 1–4

Cesarano C, Mazzeo A, Picariello A (2007) A System for Summary-document Similarity in Notary domain. Proceedings of the International Workshop on Database and Expert Systems Applications pp 254–258

Chaovalit P, 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.

Christensen HU, Ortiz-Arroyo D (2007) Applying data fusion methods to passage retrieval in QAS. In: Proceedings of the 7th international conference on Multiple classifier systems (MCS'07), Springer-Verlag, Berlin, Heidelberg, pp 82–92

Conroy J, Schlesinger J (2008) CLASSY at TAC 2008 Metrics. In: Proceedings of the Text Analysis Conference (TAC)

Cruz F, Troyano J, Ortega J, Enríquez F (2008) The Italica System at TAC 2008 Opinion Summarization Task. In: Proceedings of the Text Analysis Conference (TAC)

Dagan I, Glickman O, Magnini B (2006) The PASCAL Recognising Textual Entailment Challenge. Machine Learning Challenges Lecture Notes in Computer Science 3944:177–190

Dang HT, Owczarzak K (2009) Overview of the TAC 2009 Summarization Track. In: Proceedings of the Text Analysis Conference (TAC)

Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40(2):1–60

Dave K, Lawrence S, Pennock D (2003) Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In: Proceedings of WWW-03.

Dunlavy DM, O'Leary DP, Conroy JM, Schlesinger JD (2007) QCS: A system for querying, clustering and summarizing documents. Information Processing & Management 43(6):1588–1605

Esuli A, Sebastiani F (2006) SentiWordNet: A Publicly Available Resource for Opinion Mining. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'06), Italy, pp 417–422

Fellbaum C (1998) WordNet: An Electronic Lexical Database. The MIT Press

Ferrández O, Micol D, Muñoz R, Palomar M (2007) A Perspective-Based Approach for Solving Textual Entailment Recognition. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Association for Computational Linguistics, Prague, pp 66–71

Foote J (1999) An overview of audio information retrieval. Multimedia Systems – Special issue on audio and multimedia 7(1):2–10

Giannakopoulos G, Karkaletsis V, Vouros G (2008) Testing the Use of n-gram Graphs in Summarization Sub-tasks. In: Proceedings of the Text Analysis Conference (TAC)

Givón T (1990) Syntax: A functional-typological introduction, II, John Benjamins

Gómez JM (2007) Recuperación de Pasajes Multilingüe para la Búsqueda de Respuestas. PhD thesis, Universidad Politécnica de Valencia, Valencia, Spain

Gómez JM, Buscaldi D, Rosso P, Sanchis E (2007) Jirs: Language-independent passage retrieval system: A comparative study. In: 5th International Conference on Natural Language Processing 2006, Hyderabad, India

y Gómez MM, Pineda LV, Pérez-Coutiño MA, Soriano JMG, Arnal ES, Rosso P (2005) A Full Data-Driven System for Multiple Language Question Answering. In: Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Springer, Lecture Notes in Computer Science, vol 4022, pp 420–428

Harabagiu S, Lacatusu F (2005) Topic Themes for Multi-document Summarization. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, pp 202–209

Hatzivassiloglou V, Wiebe JM (2000) Effects of adjective orientation and gradability on sentence subjectivity. In: Proceedings of the 18th conference on

Computational linguistics - Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 299–305

He T, Chen J, Gui Z, Li F (2008) CCNU at TAC 2008: Proceeding on Using Semantic Method for Automated Summarization. In: Proceedings of the Text Analysis Conference (TAC)

Hsin-Hsi C, Chuan-Jie L (2000) A Multilingual News Summarizer. In: Proceedings of the 18th conference on Computational linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 159–165

Hu M, Liu B (2004) Mining Opinion Features in Customer Reviews. In: Proceedings of Nineteenth National Conference on Artificial Intellgience AAAI-2004

Jin F, Huang M, Zhu X (2009) A Query-specific Opinion Summarization System. In: Proceedings of 8th IEEE International Conference on Cognitive Informatics, pp 428–433

Kabadjov M, Balahur A, Boldrini E (2009) Sentiment Intensity: Is It a Good Summary Indicator? In: Proceedings of the 4th Language and Technology Conference (LTC), pp 380–384

Kaisser M, Hearst MA, Lowe JB (2008) Improving Search Results Quality by Customizing Summary Lengths. In: Proceedings of the Association for Computational Linguistics – Human Language Technologies (ACL-08: HLT), Association for Computational Linguistics, Columbus, Ohio, pp 701–709

Kan MY, Klavans JL (2002) Using Librarian Techniques in Automatic Text Summarization for Information Retrieval. In: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries (JCDL '02), ACM, New York, NY, USA, pp 36–45

Kazantseva A (2006) An Approach to Summarizing Short Stories. In: Proceedings of the Student Research Workshop at the 11th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 55–62

Kim SM, Hovy E (2004) Determining the Sentiment of Opinions. In: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 1367–1373

Koppel M, Shtrimberg I (2006) Good News or Bad News? Let the Market Decide. Computing Attitude and Affect in Text: Theory and Applications pp 297–301

Kudo T (2004) A Boosting Algorithm for Classification of Semi-Structured Text. Science And Technology (29):17–24

Kuo JJ, Chen HH (2008) Multidocument Summary Generation: Using Informative and Event Words. ACM Transactions on Asian Language Information Processing (TALIP) 7(1):1–23

Lerman K, McDonald R (2009) Contrastive Summarization: An Experiment with Consumer Reviews. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, Association for Computational Linguistics, Boulder, Colorado, pp 113–116

Lerman K, Blair-Goldensohn S, McDonald R (2009) Sentiment Summarization: Evaluating and Learning User Preferences. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Association for Computational Linguistics, Athens, Greece, pp 514–522

Lin CY (2004) ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of ACL Text Summarization Workshop, Association for Computational Linguistics, Barcelona, Spain, pp 74–81

Lin CY, Hovy E (2000) The Automated Acquisition of Topic Signatures for Text Summarization. In: Proceedings of the 18th Conference on Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 495–501

Liu B (2006) Web Data Mining. Exploring Hyperlinks, Contents and Usage Data, 1st edn. Springer

Lloret E (2011) Text Summarisation based on Human Language Technologies and its Applications. PhD thesis, University of Alicante

Lloret E, Palomar M (2009) A Gradual Combination of Features for Building Automatic Summarisation Systems. In: Proceedings of the 12th International Conference on Text, Speech and Dialogue (TSD), Springer-Verlag, Berlin, Heidelberg, pp 16–23

Lloret E, Ferrández O, Muñoz R, Palomar M (2008a) Integración del reconocimiento de la implicación textual en tareas automáticas de resúmenes de textos. Procesamiento del Lenguaje Natural (41):183–190

Lloret E, Ferrández O, Muñoz R, Palomar M (2008b) 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) in conjunction with the 10th International Conference on Enterprise Information Systems (ICEIS 2008), 12-16 June, Barcelona, Spain, pp 22–31

Lloret E, Balahur A, Palomar M, Montoyo A (2009) Towards Building a Competitive Opinion Summarization System: Challenges and Keys. In: Proceedings of the North American Chapter of the Association for Computational Linguistics. Student Research Workshop and Doctoral Consortium, pp 72–77

Lloret E, Saggion H, Palomar M (2010) Experiments on Summary-based Opinion Classification. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, Los Angeles, CA, pp 107–115

Lloret E, Romá-Ferri MT, Palomar M (2011) COMPENDIUM: A Text Summarization System for Generating Abstracts of Research Papers. In: Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB)

Luhn HP (1958) The Automatic Creation of Literature Abstracts. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press, pp 15–22

Ma L, He T, Li F, Gui Z, Chen J (2008) Query-Focused Multi-document Summarization Using Keyword Extraction. Proceedings of the 2008 International Conference on Computer Science and Software Engineering - Volume 01 pp 20–23

Mani I (2001) Automatic Summarization. John Benjamins Pub Co

Mani I, House D, Klein G, Hirschman L, Firmin T, Sundheim B (1999) The TIPSTER SUMMAC Text Summarization Evaluation. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 77–85

Manning CD, Raghavan P, Schtze H (2008) Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA

McCargar V (2005) Statistical Approaches to Automatic Text Summarization. Bulletin of the American Society for Information Science and Technology 30(4):21–25

McKeown K, Radev DR (1999) Generating Summaries of Multiple News Articles. In: Inderjeet Mani and Mark Maybury, editors, *Advances in Automatic Text Summarization*, MIT Press, pp 381–390

Mihalcea R, Ceylan H (2007) Explorations in Automatic Book Summarization. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp 380–389

Mishne G (2006) Multiple Ranking Strategies for Opinion Retrieval in Blogs. 2006 TREC Blog Track

Najork M, Heydon A (2002) Handbook of Massive Data Sets. Kluwer Academic Publishers, Norwell, MA, USA, chap High-performance Web Crawling, pp 25–45

Nenkova A, Passonneau R, McKeown K (2007) The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. ACM Transactions on Speech and Language Processing 4(2):4

Ng V, Dasgupta S, Arifin SMN (2006) Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In: Proceedings of the COLING/ACL on Main conference poster sessions, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 611–618

Ou S, Khoo CSG, Goh DH (2007) Automatic Multidocument Summarization of Research Abstracts: Design and User Evaluation. Journal of American Society for Information Science and Technology 58(10):1419–1435

Ounis I, de Rijke M, Macdonald C, Mishne G, Soboroff I (2006) Overview of the TREC-2006 Blog Track. In: Proceeddings of the 15th Text REtrieval Conference (TREC 2007)

Pang B, 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 Association for Computational Linguistics, pp 115–124

Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL 2004

Pang B, Lee L (2008) Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135

Pang B, Lee L, 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, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 79–86

Plaza L, Lloret E, Aker A (2010) Improving Automatic Image Captioning Using Text Summarization Techniques. In: Proceedings of the 13th International Conference on Text, Speech and Dialogue (TSD), Springer-Verlag, Berlin, Heidelberg, pp 165–172

Radev DR, Fan W (2000) Automatic Summarization of Search Engine Hit Lists. In: Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Association for Computational Linguistics, Morristown, NJ, USA, pp 99–109

Riloff E, Wiebe J (2003) Learning Extraction Patterns for Subjective Expressions. In: Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 105–112

Riloff E, Wiebe J, Phillips W (2005) Exploiting subjectivity classification to improve information extraction. In: Proceedings of the 20th national conference on Artificial intelligence - Volume 3, AAAI Press, pp 1106–1111, URL http://portal.acm.org/citation.cfm?id=1619499.1619511

Saggion H (2009) A Classification Algorithm for Predicting the Structure of Summaries. In: Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009), Association for Computational Linguistics, Suntec, Singapore, pp 31–38

Sakai T, Sparck-Jones K (2001) Generic Summaries for Indexing in Information Retrieval. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01), ACM, New York, NY, USA, pp 190–198

Sauper C, Barzilay R (2009) Automatically Generating Wikipedia Articles: A Structure-Aware Approach. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, Suntec, Singapore, pp 208–216

Scherer K (2005) What are Emotions? And How can they be Measured? Social Science Information 44(4):693–727

Shen D, Yang Q, Chen Z (2007) Noise Reduction through Summarization for Web-page Classification. Information Processing and Management 43(6):1735 – 1747

Soriano JMG, Buscaldi D, Asensi EB, Rosso P, Arnal ES (2005) QUASAR: The Question Answering System of the Universidad Politcnica de Valencia. In: Peters C, Gey FC, Gonzalo J, Mller H, Jones GJF, Kluck M, Magnini B, de Rijke M (eds) Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, Springer, Lecture Notes in Computer Science, vol 4022, pp 439–448

Steinberger J, Jezek K, Sloup M (2008) Web Topic Summarization. In: Proceedings of the 12th International Conference on Electronic Publishing (ELPUB, Toronto, Canada, 25-27 June, pp 322–334

Stoyanov V, Cardie C (2006) Toward Opinion Summarization: Linking the Sources. In: Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp 9–14

Strapparava C, Valitutti A (2004) WordNet-Affect: an Affective Extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, pp 1083–1086

Teufel S, Moens M (2002) Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. Computational Linguistis 28(4):409–445

Titov I, McDonald R (2008) A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In: Proceedings of Association for Computational Linguistics – Human Language Technologies (ACL-08: HLT), Association for Computational Linguistics, Columbus, Ohio, pp 308–316

Torres-Moreno JM, St-Onge PL, Gagnon M, El-Bze M, Bellot P (2009) Automatic Summarization System coupled with a Question-Answering System (QAAS). NLP News Computation and Language

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

Wilson T, Wiebe J, Hwa R (2004) Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In: Proceedings of the 19th national conference on Artifical intelligence, AAAI Press, pp 761–767

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 347–354

Witten IH, Moffat A, Bell TC (1999) Managing Gigabytes : Compressing and Indexing Documents and Images, 2nd edn. Morgan Kaufmann, San Francisco, CA

Yang K (2008) WIDIT in TREC 2008 Blog Track: Leveraging Multiple Sources of Opinion Evidence. In: Proceedings of The 17th Text REtrieval Conference

Yang XP, Liu XR (2008) Personalized Multi-document Summarization in Information Retrieval. Proceedings of International Conference on Machine Learning and Cybernetics 7:4108–4112

Yu D, Hatzivassiloglou V (2003) Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In: Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics, Stroudsburg, PA, USA, pp 129–136

Zhang W, Yu C, Meng W (2007) Opinion Retrieval from Blogs. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM, New York, NY, USA, pp 831–840

Zhao L, Wu L, Huang X (2009) Using Query Expansion in Graph-based Approach for Query-focused Multi-document Summarization. Information Processing and Management 45(1):35–41

Zhuang L, Jing F, Zhu XY (2006) Movie Review Mining and Summarization. In: Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM '06), ACM, New York, NY, USA, pp 43–50