INTERNATIONAL WORKSHOP

EVENTS IN EMERGING TEXT TYPES (eETTs)

held in conjunction with the International Conference

RANLP - 2009, 14-16 September 2009, Borovets, Bulgaria

PROCEEDINGS

Edited by
Constantin Orăsan, Laura Hasler and Corina Forăscu

Borovets, Bulgaria

17 September 2009

International Workshop

EVENTS IN EMERGING TEXT TYPES (eETTs)

PROCEEDINGS

Borovets, Bulgaria
17 September 2009

ISBN 978-954-452-011-3

Designed and Printed by INCOMA Ltd.
Shoumen, Bulgaria

Summarizing Threads in Blogs Using Opinion Polarity

Alexandra Balahur[1,2], Elena Lloret[1], Ester Boldrini[1],
Andrés Montoyo[1], Manuel Palomar[1], Patricio Martínez-Barco[1]

[1]Natural Language Processing and Information Systems Group
Dept. of Software and Computing Systems, University of Alicante
Apartado de Correos 99, E-03080, Alicante, Spain
[2] European Commission Joint Research Centre
Institute for the Protection and Security of the Citizen
Global Security and Crisis Management Unit, OPTIMA Action
Via E. Fermi, 2749, I-21027, Ispra (VA), Italy
{abalahur, elloret, eboldrini, montoyo, mpalomar, patricio}@dlsi.ua.es

Abstract

The huge amount of data available on the Web needs to be organized in order to be accessible to users in real time. This paper presents a method for summarizing subjective texts based on the strength of the opinion expressed in them. We used a corpus of blog posts and their corresponding comments (blog threads) in English, structured around five topics and we divided them according to their polarity and subsequently summarized. Despite the difficulties of real Web data, the results obtained are encouraging; an average of 79% of the summaries is considered to be comprehensible. Our work allows the user to obtain a summary of the most relevant opinions contained in the blog. This allows them to save time and be able to look for information easily, allowing more effective searches on the Web.

Keywords
Opinion Mining, Sentiment Analysis, Blog Posts, Automatic Summarization.

1. Introduction

Due to the rapid development of the Social Web, new textual genres expressing subjective content by means of emotions, feelings, sentiments, moods or opinions are growing rapidly. Nowadays, people converse frequently using many non-conventional ways of communication such as blogs, forums or reviews. As a consequence, the number of such emerging text types is growing at an exponential rate, as well as their impact on the everyday lives on millions of people.

A research for the Pew Institute [1] shows that 75,000 blogs are created per day by people all over the world, on a great variety of subjects. Thus, blogs are becoming an extremely relevant resource for different kinds of studies focused on many useful applications. This research area have become known as *sentiment analysis* or *opinion mining.* However, as there is no overall accepted definition of this task and in order to delimit our research area, the concepts of emotions, feelings, sentiments, moods and opinions need to be defined with precision.

Emotion is "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems) in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism [2, 3].

The term "feeling" points to a single component denoting the subjective experience process [4] and is therefore only a small part of an emotion.

"Moods" are less specific and intense affective phenomena, product of two dimensions - energy and tension [5].

"Sentiment" is defined in the Webster dictionary[1] as: 1 a: an attitude, thought, or judgment prompted by feeling: predilection b: a specific view or notion: opinion; 2 a: emotion b: refined feeling: delicate sensibility especially as expressed in a work of art c: emotional idealism d: a romantic or nostalgic feeling verging on sentimentality; 3 a: an idea colored by emotion b: the emotional significance of a passage or expression as distinguished from its verbal context. Finally, the term "opinion", according to the Webster Dictionary, is 1 a: a view, judgment, or appraisal formed in the mind about a particular matter b: approval, esteem; 2 a: belief stronger than impression and less strong than positive knowledge b: a generally held view; 3 a: a formal expression of judgment or advice by an expert b: the formal expression (as by a judge, court, or referee) of the legal reasons and principles upon which a legal decision is based.

As we can deduce from these definitions, affect-related concepts are similar in nature and in many cases overlap;

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

however, we can say that emotion is the super category that includes all other abovementioned concepts.

Language employed in blogs is highly heterogeneous [6]. People with different social backgrounds write them and as a consequence, they contain highly variable and unpredictable language [7]. Moreover, surveys show that they are not entirely written using an informal style; it is only employed for a small part of them and many users aim instead at a more refined style. As we can deduce, these texts offer an example of genuine and spontaneous Natural Language, providing the opportunity of challenging studies focused on solving the problems of its understanding and generation. Not less important to mention is also the fact that blogs contain frequent "copy-pastes" from news sources that are introduced to support a point of view or argument.

It is worth mentioning that those emerging texts are extremely relevant also because bloggers write whatever is on their mind about a wide range of topics [8]. In most of the cases, they aim to share their feelings about an episode of their lives, a "hot" news topic or a product, for example [9]; consequently, these corpora provide an excellent platform research on informal communications [10].

Researchers could exploit this huge amount of data for an enormous number of applications useful for companies, economic institutions, educational centers, politic parties, etc. Companies could use them to discover the customers' preferences, complaints or to monitor opinions about competitors. Economic institutions could take advantage of this information to predict and control people's attitude towards relevant economic events, as for example the present economic crisis. Furthermore, educational institutions could employ them to know and understand students' opinion about teachers, methods or didactic materials, for example. And last but not least, politic institutions or parties would use them to know people's opinion about laws, bills or to foresee elections results. On the one hand, the growing volume of subjective information available on the Web allows for better and more informed decisions of the users, but on the other hand, the quantity of data to be analyzed imposes the automation of the opinion mining process as well as other Natural Language Processing (NLP) tasks. Our research is focused on opinion summarization of blog posts about different topics. Our main purpose is to provide the user with a summary of positive and negative opinions about a specific topic. The summary will be generated in three sizes, 10%, 15% and 20%. Depending on the user profile and its needs we would offer him the size s/he needs. For example, if we work with a blog about mobile phones we would give back a short summary if the user does not have a high level of knowledge of this product; but if the user is a technician, the system would give him back a more detailed summary, because s/he would be able to understand a more technical and detailed summary. In general, this would avoid spending much of their time reading all the reviews to find what they are looking for, as the system offers them summaries of pros and cons of a topic. This would be one of the possible ways to exploit the huge amount of data the Web offers.

2. Motivation and contribution

The explosive increase in Web communication has attracted interest in technologies for automatically mining personal opinions from different kinds of Web documents, such as product reviews, blogs or forums. These technologies would benefit users who seek reviews on certain consumer products [11].

In fact, at the time of taking a decision, more and more people search for subjective information expressed on the Web on their matter of interest and base their final decision on the information found [12]. Not less important is also the fact that people interested in news and how they are reflected in the world wide opinion often use both newspaper sources, as well as blogs, in order to follow the development of news and the corresponding opinion. For this reason, we believe opinion summarization could represent a useful tool, on the one hand to help users to take decisions quickly and, on the other hand, this would also be effective to manage the huge amount of data we have.

The first contribution this paper brings is the annotation of a collection of a corpus of blog posts together with the comments given on them (threads) in English about different topics, at the level of opinion, polarity and post/comment, as well as sentence importance. We decided to select five macrotopics that are economy, science and technology, cooking, society, and sport. We obtained a total of 51 documents containing the discussion threads (original posts and the comments made on them). The average number of comments on the post is 33.

After having collected the corpus, we employed a partial version of EmotiBlog, an annotation scheme for emotion detection in non traditional textual genres [13], labeling the opinions of the different users. We decided to employ a partial version of the model to avoid noise. In fact, EmotiBlog is a fine grained model, but for the first step of our research we only need some of the elements of the traditional annotation scheme. Subsequently, we automatically classified the polarity at a sentence and also at a document level and furthermore, we proposed a method to summarize similar opinions grouped for topics. The result is a summary of positive and negative opinions, divided according to their corresponding polarity.

3. Related work

The increasing amount of data on the Web needs to be processed in order to help users who are looking for specific information. Therefore, summarization systems are becoming more and more useful because they provide shorter versions of texts, avoiding users wasting their time. Moreover, subjective information has a high presence on the Internet, by means of forums or blogs, among others. A recent application for summarization is to combine this task with Opinion Mining, in order to produce summaries of opinions on a specific topic. Regarding opinion-oriented summaries, subjective linguistic elements have to be detected and classified first, according to their polarity, and then, they have to be grouped in a coherent fragment of text in order to produce the final summary.

Opinion summarization systems that participated in the Text Analysis Conference[2] (TAC) in 2008 such as [14], [15], [16], or [17] followed these steps. However, out of the scope of the TAC competition, we can find other interesting approaches, as well. For instance, in [18] Machine Learning algorithms are used to determine which sentences should belong to a summary, after identifying possible opinion text spans. The useful features to locate opinion quotations within a text included location within the paragraph and document, and the type of words they contained. Similarly, in [19] the relevant features and opinion words with their polarity (whether a positive or a negative sentiment) are identified, and then, after detecting all valid feature-opinion pairs, a summary is produced, but focusing only in movie reviews. Normally, online reviews also contain numerical ratings that users insert when providing their personal opinions about a product or service. In [20] a Multi-Aspect Sentiment model is proposed. This statistical model uses aspect ratings to discover the corresponding topics and extract fragments of text.

Our work differs from the ones abovementioned since we take into account the posts written in real blogs, to further build a summary of the most relevant opinions contained in them, based on their polarity.

4. Corpus collection and labeling

The corpus we employed in this study is a collection of 51 blogs extracted from the Web. This is a limited dataset which allows for a preliminary study in the field; however, in our future work we would like to extend it in order to carry out a more in depth research. The blog posts are written in English and have the same structure Generally, blogs have the following organization: the authors create an initial post containing a piece of news and their opinion on it and subsequently, bloggers reply expressing their opinions about the topic. In most of the cases, commenting posts are the most subjective texts even if also in its first intervention the author can express its point of view. They can also contain multimodal information, but we decided to take into account only the text; however, the multimodal information analysis could be an interesting research for future work. In our blog corpus annotation, we indicated the *url* from which the thread was extracted it, we then included the initial annotated piece of news and the labeled user comments.

People use this new textual genre to express opinions on a wide range of topics. However, in order to delimitate our work, we were forced to select only few of them; we gave priority to the most relevant threads, that contained a large amount of posts in order to have a considerable amount of data. We chose some of the topics that we considered relevant: economy, science and technology, cooking, society and sport. Regarding its size, Table 1 shows the average and the total number of posts, of words in the news, of the number of words in posts and, finally, of words both in news and in posts.

Table 1: Corpus size

	N. Posts	N. Words for new	N. Word for post	Total words
Total	1829	72.995	226.573	299.568
Average	33.87	1351.75	4195.79	5547.55

As can be seen in Table 1, we did not work with a huge corpus. In fact, this is a work in progress.We started with a small quantity of data, but one of our objectives is to annotate more data in order to be able to use a bigger corpus and compare the results. After having collected the corpus, we labelled it using some of the EmotiBlog elements presented in Table 2.

Table 2: Annotated elements

Element	Attribute
Polarity	Positive, negative
Level	Low, medium, high
Source	name
Target	name

As we can see in Table 2, we decided to select only a few of the elements in EmotiBlog [12]; each of them has been chosen with a special purpose. Firstly, we discriminated between objective and subjective sentences, and after that, we took into consideration only the subjective sentences with the elements presented in the table. Each of the elements indicated in the table above has been selected

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

because they provide important information that is relevant to the task at hand. The polarity has the function of indicating if the opinion expressed in the sentence is positive of negative. Moreover, we labeled the data at the opinion level, choosing the level of polarity intensity between low, medium or high. Finally, we specified the source of the discourse in order to be able to detect who said what, and the target of the sentence, so as to understand the topic of the discourse. We decided not to include all the elements of EmotiBlog to avoid noise. The result of the annotation process is a gold standard which will be used to evaluate some of the aspects of the generated summaries. The subjective sentences are annotated with polarity, the level of this polarity and also with the source and the target of the discourse.

Figure 1: Example of labeling

<topic>economic situation**</topic>**

<topic2>government**</topic2>**

<topic3>banks**</topic3>**

<new> Saturday, May 9, 2009 My aim in this blog has largely been to give my best and most rational perspective on the reality of the economic situation. I have tried (and I hope) mostly succeeded in avoiding emotive and partisan viewpoints, and have tried as far as possible to see the actions of politicians as misguided. Of late, that perspective has been slipping, for the UK, the US and also for Europe.

<phenomenon gate:gateId="1" target="economic crisis" degree1="medium" category="phrase" source="Cynicus Economicus" polarity1="negative" >I think that the key turning point was the Darling budget, in which the forecasts were so optimistic as to be beyond any rational belief**</phenomenon>**…

Figure 1 is an example of annotation. We would like to stress upon the fact that we indicate more than one topic. We decided to contemplate cases of multiple topics only if they are relevant in the blog. In this case, the main topic is the economic situation, while the secondary ones are the government and banks.

After having defined the topics, the first paragraph contains objective information and thus, we do not label it; we therefore annotate the following sentence that contains subjective information. As you can see, the economic crisis is the target. Finally, the polarity of the sentence is negative, the intensity level of this polarity is medium and the author is Cynicus Economicus.

4.1 Annotation problems

During the annotation process we faced some difficulties, to which we tried proposing possible solutions.

The first obstacle we detected consisted in finding the topic of each blog. We started with the assumption that generally the title gives the idea of a topic, but , after having read the posts, we realized that the topic is not just the one included in the idea of the title. Furthermore, it is very usual that the author of the new writes about a topic, but during the discussion in the blog, people change the topic of conversation. In order to overcome these problems, we decided to insert more than one topic, given that they are relevant to the global discourse. There are also blogs where no specific topic is addressed and where people talk about many different subjects and express opinions on each of them.

5. Generating summaries from posts

In order to produce summaries from blogs, and, more specifically, from the posts about news, we used, as a core for the summarization process, the summarization approach proposed in [author's reference]. However, as this system produces generic summaries, the blog posts had to be pre-processed and classified according to their polarity before producing the final summaries. Therefore, two sub-tasks can be distinguished within the whole process: sentence polarity classification and summary generation.

5.1 Sentence polarity classification

The first step we took in our approach was to determine the opinionated sentences, assign each of them a polarity (among positive and negative) and a numerical value corresponding to the polarity strength (the higher the negative score, the more negative the sentence and similarly, the higher the positive score, the more positive the sentence). Given that we are faced with the task of classifying opinion in a general context, we employed a simple, yet efficient approach, presented in [25]. 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 [22], SentiWordNet [23], MicroWNOp [24]. Each of the employed resources were mapped to four categories, which were given different scores: positive (1), negative (-1), high positive (4) and high negative (-4). As shown in [25], these values performed better than the usual assignment of only positive (1) and negative (-1) values. First, the score of each of the blog posts was computed as sum of the values of the words 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[3] and classified the obtained sentences according to their polarity, by adding the individual scores of the affective words identified. As it has been shown in [25], some resources tend to over classify positive or negative examples. Thus, we have used the combined resources, which have proven to classify in a more balanced manner [25]. The measure of the intensity of the scores can also be used as an indication

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

of the sentence importance and can thus constitute a criterion for summarization, as shown in [16].

5.2 Summary generation

Once all subjective sentences have been classified, we grouped them according to their polarity, distinguishing between positives and negatives. It is worth mentioning that, although the polarity of all blog sentences was determined, we only took into consideration the ones belonging to the comment posts and not in the initial news post of the blogs. This was motivated by the fact that the purpose of our summaries is to contain opinions stated by the users who have already read that news and want to express their thoughts in relation to it.

One of the main problems of blogs as far as a type of document is concerned, is the big amount of noisy information they contain. This fact can affect the quality of final summaries, and in order to avoid this, we decided to run a pre-process step, removing all unnecessary information. The problem is how to determine which information is necessary and which is not. For the purpose of our experiments, we decided that the person who stated the opinion as well as the date and time the post was written would be considered as noisy information. In some particular cases, it would be interesting to keep this information so that different strategies for grouping opinions and presenting the summary could be taken into account, such as the analysis of all the opinions of the same person. At the moment, we are more interested in subjective sentences, so that we can summarize them to provide users with the main opinions about a topic. Another problem found was the difficulty in detecting noisy information from the blogs, since each one of them presents the information in different formats. For example, regarding the authors of the posts we can find fragments such as "Paul said...", "drpower said 2:05PM on 5-13-2009", "# Julie May 14, 2009", or "Adrian Eden - May 14th, 2009 at 8:43 pm PDT". To tackle this problem, we decided to analyze the set of blogs we had and detect how the unnecessary information we wanted to remove was written; as a consequence, several manual rule-based patterns could be designed to identify this information. Having all sentences without noisy information, the next step was to run the summarization approach. It is worth mentioning that the blogs may contain orthographic and grammatical errors, which may also affect the quality of the final summaries. However, we decided not to correct them in order to maintain all the features of this kind of emerging genre. This approach employs textual entailment to remove redundant information, and computes word-frequency and noun-phrases length to detect relevant sentences within a document. The output of the system is an extract, which means that the most important sentences are extracted to produce the final summary. More specific details about the features of the summarization approaches can be found in [21]. Two different summaries were produced for each blog, one with the positive opinions and one with the negative ones. Finally, as a post-processing stage, we bound together the summaries belonging to the same blog to produce the final summary. In the end, we generated 51 opinion summaries from different topics (economy, science and technology, cooking, society and sport), one corresponding to each blog of the corpus described in the previous sections.

6. Evaluation

The evaluation of summaries is a difficult task. On the one hand, automatic systems for evaluating summaries require reference summaries written by humans, and this is a very time-consuming task. Moreover, different humans would produce diverse summaries, resulting in several possible correct summaries as gold standard, making this fact another problem for the evaluation. In [26] it was shown how the result for a summary changed depending on which human summary was taken as reference for comparison with the automatic one. This problem was also presented in [27] and [28]. More recently, in [29] they stated the need of performing a more qualitative evaluation rather than a quantitative one, since summaries must contain relevant information, but at the same time, they should have an acceptable quality in order to be useful for other tasks or applications. In the DUC[4] and TAC conferences, summaries are evaluated manually taking into account several linguistic quality criteria, such as grammaticality or structure and coherence, for example. In this paper, we have adopted a similar approach for evaluating the generated summaries. We focus more on the quality of the summaries rather than on its content, since the content would depend on the specific need a user has at a particular moment; this has not been taken into consideration yet in our approach. However, for future work, it would be interesting to study and analyze how to produce different summaries depending on a user's profile. The criteria proposed for evaluating the opinion summaries are the following: redundancy, grammaticality, focus and difficulty. Redundancy measures the presence of repeated information in a summary. Grammaticality accounts for the number of spelling or grammatical errors that a summary presents. Focus evaluates whether it is possible or not to understand the topic of the summary, that is, the main subject of the text; and finally, difficulty refers to the extent to which a human can understand a summary as a whole or not. As can be seen, we took as a basis the criteria proposed in DUC and TAC conferences, except from the difficulty criteria which is non-conventional. We decided to contemplate this criterion, because it could be a method to evaluate the overall summary. For each one of them,

[4] http://www-nlpir.nist.gov/projects/duc

three different degrees of goodness were established. These were non-acceptable, understandable and acceptable. In this classification, acceptable means that the summary meets the specific criterion and therefore is good, whereas non-acceptable would mean that the summary would not be good enough with respect to a criterion. When measuring difficulty, the summaries were classified with regard to high, medium and low, being low, the better. When we evaluate the summaries with this criterion, some factors must be taken into account. The first one is the grammatical correctness; the length of the summary is another relevant element, because in fact, it is more difficult to evaluate big summaries than short ones, although longer summaries become more clear in content and understandable than short ones, as demonstrated by the results obtained. The third one is the topic. We consider as good summaries only those where the topic is clear through the text and finally, the last element is the background of the supervisor. We are convinced that evaluating a summary manually could be a very subjective task because it depends on the different backgrounds the evaluators have. The higher their level is, the clearer the summary will be.

The evaluation has been manually carried out by two potential users who, although not experts in evaluating summaries, would be very interested in having such an application to process what people think about a specific topic.

While revising the summaries, we noticed some recurrent mistakes. The first one is the punctuation; in some cases we noticed some commas missing or instead of having a comma, contain a full stop. (e.g. ‘So. One opition…’) Also, in some cases, apostrophes are missing, in examples such as ‘don’t’..

The second is that sometimes we find ‘PDTAh, yea’h, for example; this is the result of regular expressions that have not been processed correctly.

The third error is that in some cases the summaries start with a sentence containing a correference element that we cannot resolve, because the antecedent has been deleted or sentences that imply some concept previously mentioned in the original text that have not been selected.

It is also worth mentioning that some of the grammatical errors are due to users’ misspellings, for example ‘I thikn’.

Finally, we also found some void sentences, that do not contribute to the general meaning of the summary as for example, ‘I m an idiot’, ‘Just an occasional visitor’, or ‘welcome back!!!. The tables below shows the results obtained:

Table 3: results of the evaluation for 10% compression ratio

	Non Accept.	Understand	Accept
Redun.	26%	45%	29%
Gramm.	4%	22%	74%
Focus	33%	43%	24%

Table 4: results of the evaluation for 15% compression ratio

	Non Accept.	Understand	Accept
Redun.	0%	6%	94%
Gramm.	2%	27%	71%
Focus	26%	29%	45%

Table 5: results of the evaluation for 20% compression ratio

	Non Accept.	Understand	Accept
Redun.	4%	10%	86%
Gramm.	0%	55%	45%
Focus	14%	47%	39%

Table 6: results for the difficulty parameter

	High	Medium	Low
10%	35%	28%	37%
15%	18%	35%	47%
20%	8%	51%	41%

As you can see in these tables, we decided to create summaries at three different compression ratios (10%, 15% and 20%), in order to analyze the impact of the size of a summary. The compression ratio can be defined as how much shorter the summary is with respect to the original document and it can be computed dividing the length of the summary by the length of the source text [30]. The different summary sizes would allow us to draw conclusions about the length of the summary and the qualitative evaluation. Figure 2 shows an example of generated summary for the blog 29 with a compression ratio of 10 %.

Figure 2: an example of 10% ratio summary

Clothilde, I love the wallpapers!

They keep everything tasty and fresh!

Thanks a lot for the gorgeous calender desktop background.

What a great idea and beautiful photo.

I've just started recreating some of the easier and more attainable recipes.

Another lovely calendar! Clotilde, have you discontinued your "Bonjour mois" newsletter?

I'm terribly late this month but was enjoying the cheese so much that I just forgot! The peas are another winner of course.

My only quibble would be about the name.

The figure above is an example of automatic summary. As it can be seen, only opinions have been considered and these are presented grouped into positives, on the one hand and negatives, on the other. We considered it as good due to the fact that there are no objectives or useless sentences. The system presents subjective sentences with an emotional charge, and as a consequence this summary meets our purposes.

As you can see, the first part of the summary is composed by positive opinions and the last part by negative ones. The negative part starts with the sentence "My only quibble would be about the name". You could notice some spelling mistakes, which are contained in the initial blog posts Therefore, we consider as necessary to include in our system a spelling corrector in order to avoid such mistakes.

6.1 Discussion

Analysing the results obtained, we can draw a set of interesting conclusions. As far as the grammaticality criterion is concerned, the results show a decrease of grammaticality errors as the size of the summary lowers. We can see that the number of acceptable summaries varies from 74% to 45%, for a compression ratio of 20% and 10%, respectively. This is obvious, because the longer the summary, the more chances are for it to have orthographic or grammatical errors. Due to the informal language used in blogs, we thought *a priori* that summaries would contain many spelling mistakes. Contrary to this thought, generated summaries are quite well-written, only 4% of them, at most, being non-acceptable. Another important fact that can be inferred from the results is related to how the summaries deal with the topic. According to the percentages shown in the tables presented previously, the number of summaries that have correctly identified the topic and have therefore been evaluated as acceptable, changes considerably with respect to the different summary sizes, increasing when we change from 10% to 15%, but decreasing when changing from 15% to 20%. However, as a general trend, we can see that when taking into account the number of summaries that have not performed correctly in the focus parameter, there is a decreasing trend, reducing the incorrect summaries from 33% to 14%. This means that for longer summaries, the topic may be stated along the summary, although not necessarily in the beginning of it, whereas for shorter summaries, there is no such flexibility, and as a consequence, if the topic does not appear in the beginning, the most probable thing is that it does not appear in the summary at all. Finally, regarding redundancy, results are not conclusive, since they experiment variations in size and degree of goodness, so we cannot establish any trend. What can be seen from the results is that the summaries of 20% size obtain the best results on average over the rest of the size experimented with. This is due to the fact that this compression ratio achieves higher percentage (for the understand and accept degrees of goodness) in two (grammaticality and focus) out of the three criteria proposed. Only the 15 % compression ratio summaries obtained better results in the redundancy criterion.

On the other hand, as far as the difficulty criteria concerned, results are also encouraging. According to the evaluation performed, the longer the summaries, the easier they are to understand in general. Grouping the percentages of summaries, we obtained that 65%, 82% and 92% of the summaries of size 10%, 15% and 20%, respectively, have, either medium or low level of difficulty, which give us an idea of they could be understand as a whole without serious difficulties. Again, for this criterion, the 20% summaries achieve the best results; this has also been proven by previous researches, which demonstrated that this compression ratio is more suitable for an acceptable quality of summaries [31]. It is worth mentioning that this criterion is rather subjective and depends to a large extent on different factors, such as the knowledge the person who reads the summaries, the number of grammatical errors the text contain, or the connectedness of the sentences. Moreover, it is reasonable to think that long summaries can be more difficult to understand, but our experiments show that is it actually the other way around, because longer summaries may contain more information than short ones, which allows the user to have more awareness of the content and what the summary is about.

7. Conclusion

In this paper we collected a corpus of blogs together with the comments given on them. This is an English corpus about five topics: economy, science and technology, cooking, society, and sport.

After having collected the corpus, we labeled it using a partial version of *EmotiBlog* [12], an annotation scheme for non-traditional textual genres. Furthermore, we automatically classified the polarity at sentence and also at a document level. Finally, we proposed a method for automatic summarization of similar opinions grouped for topics. The result is a summary of positive and negative opinions, divided according to their corresponding polarity. We decided to generate three different ratio summaries: 10%, 15% and 25%. In fact depending on the user's profile a different size of summary could be more convenient that another one.

We evaluated summaries taking into consideration different parameters: redundancy, grammaticality, focus and difficulty, obtaining encouraging results.

There is no doubt about the fact that opinion summarization is a challenging task. For this reason, as future work we would like to improve our method in order to obtain better summaries. The first step would consist in evaluating our work using summaries made by humans; this is a very time consuming task, however it is fundamental in order to assure the quality of our results. Furthermore, we would like to integrate some correference resolution systems that could improve the quality of the language of summaries; we have some cases of noun repetitions, or in other cases, there is a sentence with a pronoun and we do not have the antecedent in the text. Another interesting challenge would be the automatic topic detection throughout the thread. Finally, we would also like to employ our techniques to other languages, such as Spanish and Italian.

8. Acknowledgements

The research of Elena Lloret has been supported by the FPI grant (BES-2007-16268) from the Spanish Ministry of Science and Innovation, under the project TEXT-MESS (TIN2006-15265-C06-01). The research has been also partially founded by the European Commission under FP6project QALL-ME number IST-033860.

9. References

[1] P. Bo and L .Lee. Opinion Mining and sentiment analysis. Foundations and trends R. In Information Retrieval Vol. 2, Nos. 1-2 (2008) 1- 135, 2008.inguistics, and Speech Recognition. Prentice Hall, New Jersey, 2000.

[2] K.R. Scherer. Toward a Dynamic Theory of Emotion: The Component Process Model of Affective States. Geneva Studies in Emotion and Communication 1: 1–98. 1987.

[3] K.R. Scherer. Appraisal Considered as a Process of Multi-Level Sequential Checking, in K.R. Scherer, A. Schorr and T. Johnstone (eds) Appraisal Processes in Emotion: Theory,Methods, Research, pp. 92–120.New York and Oxford: Oxford University Press. 2001.

[4] K.R. Scherer. What are emotions? And how can they be measured?" Social Science Information. 44(4), 693–727. 2005.

[5] Robert E. Thayer. Calm Energy: How People Regulate Mood With Food and Exercise. Oxford University Press (New York, NY). 2001. ISBN 0-19-513189-4.

[6] M. Tavosanis. Linguistic features of Italian blogs: literary language. New Text. Wikis and blogs and other dynamic text sources, pp 11-15, Trento, Vol. 1, 2006.

[7] C, S. Corvalán. Sociolingüística y gramática del español. Washington DC: Georgetown University press, 2001.

[8] H. Qu, A. La Pietra and S. Poon. Classifying Blogs Using NLP: Challenges and Pitfalls. AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs 2006.

[9] C. Yang, K. Lin, H.-H. Chen. Emotion Classification Using Web Blog Corpora. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence Pages 275-278 Year of Publication: 2007 ISBN: 0-7695-3026-5.

[10] L.E. Holzman, W.M. Pottenger. Classification of Emotions in Internet Chat: An Application of Machine Learning Using Speech Phonemes. 2003. Lehigh CSE 2003 Technical Reports.

[11] N. Kobayashi, K. Inui, Y. Matsumoto. Extracting Aspect-Evaluation and Aspect-Of Relations in Opinion Mining. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 1065-1074. 2007.

[12] E. Boldrini, A. Balahur, P. Martínez-Barco, A. Montoyo. EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genres. In Proceedings of the 5th International Conference on Data Mining. Las Vegas, Nevada, USA. 2009.

[13] A. Balahur, E. Boldrini, A. Montoyo, P. Martínez-Barco. Fact versus Opinion Questions Classification and Answering: Challenges and Keys. In ICAI'09 - The 2009 International Conference on Artificial Intelligence. Las Vegas, Nevada, USA. 2009.

[14] J. Conroy and S. Chlesinger. 2008. Classy at TAC 2008 metrics. In Proceedings of the Text Analysis Conference (TAC).

[15] T. He, J. Chen Z. Gui and F. Li.. Ccnu at TAC 2008: Proceeding on using semantic method for automated summarization. In Proceedings of the Text Analysis Conference (TAC). 2008.

[16] A. Balahur, E. Lloret, O. Ferrández, A. Montoyo, M. Palomar and R. Muñoz. The dlsiuaes team's participation in the TAC 2008 tracks. In Proceedings of the Text Analysis Conference (TAC). 2008.

[17] A. Bossard, M. Généreux and T. Poibeau. Description of the lipn systems at TAC 2008: Summarizing information and opinions. In Proceedings of the Text Analysis Conference (TAC). 2008.

[18] P. Beineke, T. Hastie C. manning and S. Vaithyanathan. An exploration of sentiment summarization. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, J. G. Shanahan, J. Wiebe, and Y. Qu, Eds. Stanford, US. 2004.

[19] L. Zhang, F. Jing, X-Y. Zhu. Movie review mining and summarization. In CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management. 43–50. 2006.

[20] I. Titov and R. Mc Donald. A joint model of text and aspect ratings for sentiment summarization. In Proceedings of ACL-08: HLT. Columbus, Ohio, 308–316. 2008.

[21] E. Lloret, M. Palomar: 2009. A Gradual Combination of features for Building Automatic Summarisation Systems. Lecture Notes in Computer Science. 12th International Conference on Text, Speech and Dialogue.

[22] C. Strapparava, A. Valitutti. WordNet-Affect: an affective extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004, pp. 1083-1086. 2004.

[23] A. Esuli and F. Sebastiani. SentiWordNet: A Publicly Available Resource for Opinion Mining. In Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2006, Italy. 2006.

[24] S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli and G. Gandini. Language resources and linguistic theory: Typology, second language acquisition, English linguistics, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano, IT. 2007.

[25] A. Balahur, R. Steinberger, E. van der Goot and B. Pouliquen. Opinion Mining from Newspaper Quotations. In Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content, 2009 IEEE/WIC/ACM International Conference on Web Intelligence held in conjunction with IAT'09, September 2009, Milan, Italy.

[26] R. Donaway, K. W. Drummey and L. A. Mather. A comparison of rankings produced by summarization evaluation measures. In Proceedings of NAACL-ANLP 2000 Workshop on Automatic Summarization. 2008.

[27] I. Mani. Summarization evaluation: An overview. In Proceedings of the North American chapter of the Association for Computational

Linguistics (NAACL). Workshop on Automatic Summarization. 2001.

[28] A. Nenkova. Summarization evaluation for text and speech: issues and approaches. In INTERSPEECH-2006, paper 2079-Wed1WeS.1. 2006.

[29] J. M. Conroy and H. T. Dang. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Coling 2008 Organizing Committee, Manchester, UK, 2008.

[30] E.H. Hovy and C.-Y. Lin. Automated Text Summarization in SUMMARIST. In I. Mani and M. Maybury (eds), Advances in Automatic Text Summarization. MIT Press, 81-94. 1999.

[31] A.H. Morris, G. M. Kasper and D. A. Adams. The effect and limitation of automated text condensing on reading comprehension performance. Information Systems Research, Vol. 3 (1) 17-35, 1992.