Towards Building a Competitive Opinion Summarization System: Challenges and Keys

Elena Lloret*, Alexandra Balahur, Manuel Palomar and Andrés Montoyo
Department of Software and Computing Systems
University of Alicante
Apartado de Correos 99, E-03080, Alicante, Spain
{elloret, abalahur, mpalomar, montoyo}@dlsi.ua.es

Abstract

This paper presents an overview of our participation in the TAC 2008 Opinion Pilot Summarization task, as well as the proposed and evaluated post-competition improvements. We first describe our opinion summarization system and the results obtained. Further on, we identify the system's weak points and suggest several improvements, focused both on information content, as well as linguistic and readability aspects. We obtain encouraging results, especially as far as F-measure is concerned, outperforming the competition results by approximately 80%.

1 Introduction

The Opinion Summarization Pilot (OSP) task within the TAC 2008 competition consisted in generating summaries from answers to opinion questions retrieved from blogs (the Blog06[1] collection). The questions were organized around 25 targets – persons, events, organizations etc. Additionally, a set of text snippets that contained the answers to the questions were provided by the organizers, their use being optional. An example of target, question and provided snippet is given in Figure 1.

Target : George Clooney
Question: Why do people like George Clooney?
Snippet 1: 1050 BLOG06-20060125-015-0025581509 he is a great actor

Figure 1. *Examples of target, question and snippet*

The techniques employed by the participants were mainly based on the already existing summarization systems. While most participants added new features (sentiment, pos/neg sentiment, pos/neg opinion) to account for the presence of positive opinions or negative ones - CLASSY (Conroy and Schlessinger, 2008); CCNU (He et al.,2008); LIPN (Bossard et al., 2008); IIITSum08 (Varma et al., 2008) -, efficient methods were proposed focusing on the retrieval and filtering stage, based on polarity – DLSIUAES (Balahur et al., 2008) - or on separating information rich clauses - italica (Cruz et al., 2008). In general, previous work in opinion mining includes document level sentiment classification using supervised (Chaovalit and Zhou, 2005) and unsupervised methods (Turney, 2002), machine learning techniques and sentiment classification considering rating scales (Pang, Lee and Vaithyanathan, 2002), and scoring of features (Dave, Lawrence and Pennock, 2003). Other research has been conducted in analysing sentiment at a sentence level using bootstrapping techniques (Riloff and Wiebe, 2003), finding strength of opinions (Wilson, Wiebe and Hwa, 2004), summing up orientations of opinion words in a sentence (Kim and Hovy, 2004), and identifying opinion holders (Stoyanov and Cardie, 2006). Finally, fine grained, feature-based opinion summarization is defined in (Hu and Liu, 2004).

2 Opinion Summarization System

In order to tackle the OSP task, we considered the use of two different methods for opinion mining and summarization, differing mainly with respect to the use of the optional text snippets provided. Our first approach (the Snippet-driven Approach)

*Elena Lloret is funded by the FPI program (BES-2007-16268) from the Spanish Ministry of Science and Innovation, under the project TEXT-MESS (TIN-2006-15265)
[1]http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html

Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium, pages 72–77,
Boulder, Colorado, June 2009.

used these snippets, whereas the second one (Blog-driven Approach) found the answers directly in the corresponding blogs. A general overview of the system's architecture is shown in Figure 2, where three main parts can be distinguished: the question processing stage, the snippets processing stage (only carried out for the first approach), and the final summary generation module. Next, the main steps involved in each process will be explained in more detail.

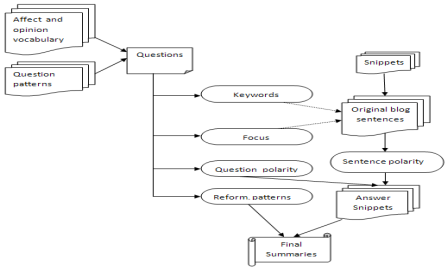

Figure 2. *System architecture*

The first step was to determine the polarity of each question, extract the keywords from each of them and finally, build some patterns of reformulation. The latter were defined in order to give the final summary an abstract nature, rather than a simple joining of sentences. The polarity of the question was determined using a set of created patterns, whose goal was to extract for further classification the nouns, verbs, adverbs or adjectives indicating some kind of polarity (positive or negative). These extracted words, together with their determiners, were classified using the emotions lists in WordNet Affect (Strapparava and Valitutti, 2005), jointly with the emotions lists of attitudes, triggers resource (Balahur and Montoyo, 2008 [1]), four created lists of attitudes, expressing criticism, support, admiration and rejection and two categories for value (good and bad), taking for the opinion mining systems in (Balahur and Montoyo, 2008 [2]). Moreover, the focus of each question was automatically extracted using the Freeling[2] Named Entity Recognizer module. This information was used to determine whether or not all the questions within the same topic had the same focus, as well as be able to decide later on which text snippet belonged to which question.

Regarding the given text snippets, we also computed their polarity and their focus. The polarity was calculated as a vector similarity between the snippets and vectors constructed from the list of sentences contained in the ISEAR corpus (Scherer and Wallbot, 1997), WordNet Affect emotion lists of anger, sadness, disgust and joy and the emotion triggers resource, using Pedersen's Text Similarity Package.[3]

Concerning the blogs, our opinion mining and summarization system is focused only on plain text; therefore, as pre processing stage, we removed all unnecessary tags and irrelevant information, such as links, images etc. Further on, we split the remaining text into individual sentences. A matching between blogs' sentences and text snippets was performed so that a preliminary set of potential meaningful sentences was recorded for further processing. To achieve this, snippets not literally contained in the blogs were tokenized and stemmed using Porter's Stemmer,[4] and stop words were removed in order to find the most similar possible sentence associated with it. Subsequently, by means of the same Pedersen Text Similarity Package as for computing the snippets' polarity, we computed the similarity between the given snippets and this created set of potential sentences. We extracted the complete blog sentences to which each snippet was related. Further on, we extracted the focus for each blog phrase sentence as well. Then, we filtered redundant sentences using a naïve similarity based approach. Once we obtained the possible answers, we used Minipar[5] to filter out incomplete sentences.

Having computed the polarity for the questions and snippets, and set out the final set of sentences to produce the summary, we bound each sentence to its corresponding question, and we grouped all sentences which were related to the same question together, so that we could generate the language for this group, according to the patterns of reformulation previously mentioned. Finally, the speech style was changed to an impersonal one, in order to avoid directly expressed opinion sentences. A POS-tagger tool (TreeTagger[6]) was used to identify third person verbs and change them to a neutral style. A set of rules to identify

[2] http://garraf.epsevg.upc.es/freeling/

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

[4] http://tartarus.org/~martin/PorterStemmer/

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

[6] http://www.ims.uni-tuttgart.de/projekte/corplex/TreeTagger/

pronouns was created, and they were also changed to the more general pronoun "they" and its corresponding forms, to avoid personal opinions.

3 Evaluation

Table 1 shows the final results obtained by our approaches in the TAC 2008 Opinion Pilot (the rank among the 36 participating systems is shown in brackets for each evaluation measure). Both of our approaches were totally automatic, and the only difference between them was the use of the given snippets in the first one (A1) and not in the second (A2). The column numbers stand for the following average scores: summarizerID (1); pyramid F-score (Beta=1) (2), grammaticality (3); non-redundancy (4); structure/coherence (including focus and referential clarity) (5); overall fluency/readability (6); overall responsiveness (7).

1	2	3	4	5	6	7
A1	0.357 (7)	4.727 (8)	5.364 (28)	3.409 (4)	3.636 (16)	5.045 (5)
A2	0.155 (23)	3.545 (36)	4.364 (36)	3.091 (13)	2.636 (36)	2.227 (28)

Table 1. *Evaluation results*

As it can be noticed from Table 1, our system performed well regarding F-measure, the first run being classified 7th among the 36 evaluated. As far as the structure and coherence are concerned, the results were also good, placing the first approach in the fourth. Also worth mentioning is the good performance obtained regarding the overall responsiveness, where A1 ranked 5th. Generally speaking, the results for A1 showed well-balanced among all the criteria evaluated, except for non redundancy and grammaticality. For the second approach, results were not as good, due to the difficulty in selecting the appropriate opinion blog sentence by only taking into account the keywords of the question.

4 Post-competition tests, experiments and improvements

When an exhaustive examination of the nuggets used for evaluating the summaries was done, we found some problems that are worth mentioning.

a) Some nuggets with high score did not exist in the snippet list (e.g. "*When buying from CARMAX, got a better than blue book trade-in on old car"* (0.9)).
b) Some nuggets for the same target express the same idea, despite their not being identical (e.g. "*NAFTA needs to be renegotiated to protect Canadian sovereignty"* and "*Green Party: Renegotiate NAFTA to protect Canadian Sovereignty"*).
c) The meaning of one nugget can be deduced from another's (e.g. "*reasonably healthy food" and "sandwiches are healthy"*).
d) Some nuggets are not very clear in meaning (e.g. "*hot", "fun").*
e) A snippet can be covered by several nuggets (e.g. both nuggets "*it is an honest book"* and *"it is a great book"* correspond to the same snippet *"It was such a great book- honest and hard to read (content not language difficulty)").*

On the other hand, regarding the use of the optional snippets, the main problem to address is to remove redundancy, because many of them are repeated for the same target, and we have to determine which snippet represents better the idea for the final summary, in order to avoid noisy irrelevant information.

4.1 Measuring the Performance of a Generic Summarization System

Several participants in the TAC 2008 edition performed the OSP task by using generic summarization systems. Most were adjusted by integrating an opinion classifier module so that the task could be fulfilled, but some were not (Bossard et al., 2008), (Hendrickx and Bosma, 2008). This fact made us realize that a generic summarizer could be used to achieve this task. We wanted to analyze the effects of such a kind of summarizer to produce opinion summaries. We followed the approach described in (Lloret et al., 2008). The main idea employed is to score sentences of a document with regard to the word frequency count (WF), which can be combined with a Textual Entailment (TE) module.

Although the first approach suggested for opinion summarization obtained much better results in the evaluation than the second one (see Section 3.1), we decided to run the generic system over both approaches, with and without applying TE, to

provide a more extent analysis and conclusions. After preprocessing the blogs and having all the possible candidate sentences grouped together, we considered these as the input for the generic summarizer. The goal of these experiments was to determine whether the techniques used for a generic summarizer would have a positive influence in selecting the main relevant information to become part of the final summary.

4.2 Results and Discussion

We re-evaluated the summaries generated by the generic system following the nuggets' list provided by the TAC 2008 organization, and counting manually the number of nuggets that were covered in the summaries. This was a tedious task, but it could not be automatically performed because of the fact that many of the provided nuggets were not found in the original blog collection. After the manual matching of nuggets and sentences, we computed the average Recall, Precision and F-measure (Beta =1) in the same way as in the TAC 2008 was done, according to the number and weight of the nuggets that were also covered in the summary. Each nugget had a weight ranging from 0 to 1 reflecting its importance, and it was counted only once, even though the information was repeated within the summary.

The average for each value was calculated taking into account the results for all the summaries in each approach. Unfortunately, we could not measure criteria such as readability or coherence as they were manually evaluated by human experts.

Table 2 points out the results for all the approaches reported. We have also considered the results derived from our participation in the TAC 2008 conference (OpSum-1 and OpSum-2), in order to analyze whether they have been improved or not. From these results it can be stated that the TE module in conjunction with the WF counts, have been very appropriate in selecting the most important information of a document. Although it can be thought that applying TE can remove some meaningful sentences which contained important information, results show the opposite. It benefits the Precision value, because a shorter summary contains greater ratio of relevant information. On the other hand, taking into consideration the F-measure value only, it can be seen that the approach combining TE and WF, for the sentences in the first approach, has beaten significantly the best F-measure result among the participants of TAC 2008 (please see Table 3), increasing its performance by 20% (with respect to WF only), and improving by approximately 80% with respect to our first approach submitted to TAC 2008.

However, a simple generic summarization system like the one we have used here is not enough to produce opinion oriented summaries, since semantic coherence given by the grouping of positive and negative opinions is not taken into account. Therefore, the opinion classification stage must be added in the same manner as used in the competition.

SYSTEM	RECALL	PRECISION	F-MEASURE
OpSum-1	0.592	0.272	0.357
OpSum-2	0.251	0.141	0.155
WF-1	**0.705**	0.392	0.486
TE+WF -1	0.684	**0.630**	**0.639**
WF -2	0.322	0.234	0.241
TE+WF-2	0.292	0.282	0.262

Table 2. *Comparison of the results*

4.3 Improving the quality of summaries

In the evaluation performed by the TAC organization, a manual quality evaluation was also carried out. In this evaluation the important aspects were grammaticality, non-redundancy, structure and coherence, readability, and overall responsiveness. Although our participating systems obtained good F-measure values, in other scores, especially in grammaticality and non-redundancy, the results achieved were very low. Focusing all our efforts in improving the first approach, OpSum-1, non-redundancy and grammaticality verification had to be performed. In this approach, we wanted to test how much of the redundant information would be possible to remove by using a Textual Entailment system similar to (Iftene and Balahur-Dobrescu, 2007), without it affecting the quality of the remaining data. As input for the TE system, we considered the snippets retrieved from the original blog posts. We applied the entailment verification on each of the possible pairs, taking in turn all snippets as Text and Hypothesis with all other snippets as Hypothesis and Text, respectively. Thus, as output, we obtained the list of snippets from which we eliminated those that

are entailed by any of the other snippets. We further eliminated those snippets which had a high entailment score with any of the remaining snippets.

SYSTEM	F-MEASURE
Best system	0.534
Second best system	0.490
OpSum-1 + TE	0.530
OpSum-1	0.357

Table 3. *F-measure results after improving the system*

Table 3 shows that applying TE before generating the final summary leads to very good results increasing the F-measure by 48.50% with respect to the original first approach. Moreover, it can be seen form Table 3 that our improved approach would have ranked in the second place among all the participants, regarding F-measure. The main problem with this approach is the long processing time. We can apply Textual Entailment in the manner described within the generic summarization system presented, successively testing the relation as Snippet1 entails Snippet2?, Snippet1+Snippet2 entails Snippet3? and so on. The problem then becomes the fact that this approach is random, since different snippets come from different sources, so there is no order among them. Further on, we have seen that many problems arise from the fact that extracting information from blogs introduces a lot of noise. In many cases, we had examples such as:

At 4:00 PM John said Starbucks coffee tastes great

John said Starbucks coffee tastes great, always get one when reading New York Times.

To the final summary, the important information that should be added is *"Starbucks coffee tastes great"*. Our TE system contains a rule specifying that the existence or not of a Named Entity in the hypothesis and its not being mentioned in the text leads to the decision of "NO" entailment. For the example given, both snippets are maintained, although they contain the same data.

Another issue to be addressed is the extra information contained in final summaries that is not scored as nugget. As we have seen from our data, much of this information is also valid and correctly answers the questions. Therefore, what methods can be employed to give more weight to some and penalize others automatically?

Regarding the grammaticality criteria, once we had a summary generated we used the module Language Tool[7] as a post-processing step. The errors that we needed correcting included the number matching between nouns and determiners as well as among subject and predicate, upper case for sentence start, repeated words or punctuation marks and lack of punctuation marks. The rules present in the module and that we "switched off", due to the fact that they produced more errors, were those concerning the limit in the number of consecutive nouns and the need for an article before a noun (since it always seemed to want to correct *"Vista"* for *"the Vista"* a.o.). We evaluated by observing the mistakes that the texts contained, and counting the number of remaining or introduced errors in the output. The results obtained can be seen in Table 4.

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

Table 4. *Grammaticality analysis*

The greatest problem encountered was the fact that bigrams are not detected and agreement is not made in cases in which the noun does not appear exactly after the determiner. All in all, using this module, the grammaticality of our texts was greatly improved.

5 Conclusions and future work

The Opinion Pilot in the TAC 2008 competition was a difficult task, involving the development of systems including components for QA, IR, polarity classification and summarization. Our contribution presented in this paper resides in proposing an opinion mining and summarization method using different approaches and resources, evaluating each of them in turn. We have shown that using a generic summarization system, we obtain 80% improvement over the results obtained in the competition, with coherence being maintained by using the same polarity classification mechanisms.

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

Using redundancy removal with TE, as opposed to our initial polarity strength based sentence filtering improved the system performance by almost 50%.
Finally, we showed that grammaticality can be checked and improved using an independent solution given by Language Tool.
Further work includes the improvement of the polarity classification component by using machine learning over annotated corpora and other techniques, such as anaphora resolution. As we could see, the well functioning of this component ensures logic, structure and coherence to the produced summaries. Moreover, we plan to study the manner in which opinion sentences of blogs/bloggers can be coherently combined.

References

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

Balahur, A. and Montoyo, A. [1]. An Incremental Multilingual Approach to Forming a Culture Dependent Emotion Triggers Database. In Proceedings of the 8th International Conference on Terminology and Knowledge Engineering, 2008.

Balahur, A. and Montoyo, A. [2]. Multilingual Feature--driven Opinion Mining and Summarization from Customer Reviews. In Lecture Notes in Computer Science 5039, pg. 345-346.

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

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.

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

Cui, H., Mittal, V., Datar, M. 2006. Comparative Experiments on Sentiment Classification for Online Product Reviews. In Proceedings of the 21st National Conference on Artificial Intelligence AAAI 2006.

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

Lloret, E., Ferrández, O., Muñoz, R. and Palomar, M. 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), pages 22–31, 2008.

Gamon, M., Aue, S., Corston-Oliver, S., Ringger, E. 2005. Mining Customer Opinions from Free Text. Lecture Notes in Computer Science.

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

Hendrickx, I. and Bosma, W.. Using coreference links and sentence compression in graph-based summarization. In Proceedings of the Text Analysis Conference (TAC), 2008.

Hu, M., Liu, B. 2004. Mining Opinion Features in Customer Reviews. In Proceedings of 19th National Conference on Artificial Intelligence AAAI.

Iftene, A., Balahur-Dobrescu, A. Hypothesis Transformation and Semantic Variability Rules for Recognizing Textual Entailment. In Proceedings of the ACL 2007 Workshop on Textual Entailment and Paraphrasis, 2007.

Kim, S.M., Hovy, E. 2004. Determining the Sentiment of Opinions. In Proceedings of COLING 2004.

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.

Riloff, E., Wiebe, J. 2003 Learning Extraction Patterns for Subjective Expressions. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing.

Scherer, K. and Wallbott, H.G. The ISEAR Questionnaire and Codebook, 1997.

Stoyanov, V., Cardie, C. 2006. Toward Opinion Summarization: Linking the Sources. In: COLING-ACL 2006 Workshop on Sentiment and Subjectivity in Text.

Strapparava, C. and Valitutti, A. "WordNet-Affect: an affective extension of WordNet". In Proceedings ofthe 4th International Conference on Language Resources and Evaluation, 2004, pp. 1083-1086.

Turney, P., 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the ACL

Varma, V., Pingali, P., Katragadda, R., Krisha, S., Ganesh, S., Sarvabhotla, K., Garapati, H., Gopisetty, H.,, Reddy, V.B., Bysani, P., Bharadwaj, R. IIT Hyderabad at TAC 2008. In Proceedings of the Text Analysis Conference (TAC), 2008.

Wilson, T., Wiebe, J., Hwa, R. 2004. Just how mad are you? Finding strong and weak opinion clauses. In: Proceedings of AAAI 2004.