Going Beyond Traditional QA Systems: Challenges and Keys in Opinion Question Answering

Alexandra Balahur
Dept. of Software and Computing Systems
University of Alicante
abalahur@dlsi.ua.es

Ester Boldrini
Dept. of Software and Computing Systems
University of Alicante
eboldrini@dlsi.ua.es

Andrés Montoyo
Dept. of Software and Computing Systems
University of Alicante
montoyo@dlsi.ua.es

Patricio Martínez-Barco
Dept. of Software and Computing Systems
University of Alicante
patricio@dlsi.ua.es

Abstract

The treatment of factual data has been widely studied in different areas of Natural Language Processing (NLP). However, processing subjective information still poses important challenges. This paper presents research aimed at assessing techniques that have been suggested as appropriate in the context of subjective - Opinion Question Answering (OQA). We evaluate the performance of an OQA with these new components and propose methods to optimally tackle the issues encountered. We assess the impact of including additional resources and processes with the purpose of improving the system performance on two distinct blog datasets. The improvements obtained for the different combination of tools are statistically significant. We thus conclude that the proposed approach is adequate for the OQA task, offering a good strategy to deal with opinionated questions.

1 Introduction

The State of the Blogosphere 2009 survey published by Technorati [1] concludes that in the past years the blogosphere has gained a high influence on a high variety of topics, ranging from cooking and gardening, to economics, politics and scientific achievements. The development of the Social Web and the new communication frameworks also influenced the way information is transmitted through communities. Blogs are part of the so-called new textual genres. They have distinctive features when compared to the traditional ones, such as newspaper articles. Blog language contains formal and informal expressions, and other elements, as repeated punctuation or emoticons (used to stress upon different text elements). With the growth in the content of the blogosphere, the quantity of subjective data of the Web is increasing exponentially (Cui et al., 2006). As it is being updated in real-time, this data becomes a source of timely information on many topics, exploitable by different applications. In order to properly manage the content of this subjective information, its processing must be automated. The NLP task, which deals with the classification of opinionated content is called Sentiment Analysis (SA). Research in this field aims at discovering appropriate mechanisms to properly retrieve, extract and classify opinions expressed in text. While techniques to retrieve objective information have been widely studied, implemented and evaluated, opinion-related tasks still represent an important challenge. As a consequence, the aim of our research is to study, implement and evaluate appropriate methods for the task of Question Answering (QA) in the opinion treatment framework.

2 Motivation and Contribution

Research in opinion-related tasks gained importance in the past years. However, there are still many aspects that require analysis and im-

[1] http://technorati.com/

Coling 2010: Poster Volume, pages 27–35,
Beijing, August 2010

provement, especially for approaches that combine SA with other NLP tasks such as QA or automatic summarization. The TAC 2008 Opinion Pilot task and the subsequent research performed on the competition data have demonstrated that answering opinionated questions and summarizing subjective information are significantly different from the equivalent tasks in the same context, but dealing with factual data. This finding was confirmed by the recent work by (Kabadjov et al., 2009). The first motivation of our work is the need to detect and explore the challenges raised by opinion QA (OQA), as compared to factual QA. To this aim, we analyze the improvements that can be brought at the different steps of the OQA process: *question treatment* (identification of expected polarity – EPT, expected source – ES and expected target –ET-), *opinion retrieval* (at the level of one and three-sentences long snippets, using topic-related words or using paraphrases), *opinion analysis* (using topic detection and anaphora resolution). This preliminary research is motivated by the conclusions drawn by previous studies (Balahur et al., 2009). Our purpose is to verify if the inclusion of new elements and methods - source and target detection (using semantic role labeling (SRL)), topic detection (using Latent Semantic Analysis), paraphrasing and joint topic-sentiment analysis (classification of the opinion expressed only in sentences related to the topic), followed by anaphora resolution (using a system whose performance is not optimal), affects the results of the system and how. Our contribution to this respect is the identification of the challenges related to OQA compared to traditional QA. A further contribution consists in adding the appropriate methods, tools and resources to resolve the identified challenges. With the purpose of testing the effect of each tool, resource and technique, we carry out a separate and a global evaluation. An additional motivation of our work is the fact that although previous approaches showed that opinion questions have longer answers than factual ones, the research done in OQA so far has only considered a sentence-level approach. Another contribution this paper brings is the retrieval at 1 and 3-sentence level and the retrieval based on similarity to query paraphrases enriched with topic-related words). We believe retrieving longer text could cause additional problems such as redundancy, coreference and temporal expressions or the need to apply contextual information. Paraphrasing, on the other hand, had account for language variability in a more robust manner; however, the paraphrase collections that are available at the moment are known to be noisy. The following sections are structured as follows: Section 3 presents the related work in the field and the competitions organized for systems tackling the OQA task. In Section 4 we describe the corpora used for the experiments we carried out and the set of questions asked over each of them. Section 5 presents the experimental settings and the different system configurations we assessed. Section 6 shows the results of the evaluations, discusses the improvements and drops in performance using different configurations. We finally conclude on our approaches in Section 7, proposing the lines for future work.

3 Related Work

QA can be defined as the task in which given a set of questions and a collection of documents, an automatic NLP system is employed to retrieve the answer to the queries in Natural Language (NL). Research focused on building factoid QA systems has a long tradition; however, it is only recently that researchers have started to focus on the development of OQA systems. (Stoyanov et al., 2005) and (Pustejovsky and Wiebe, 2006) studied the peculiarities of opinion questions. (Cardie et al., 2003) employed opinion summarization to support a Multi-Perspective QA system, aiming at identifying the opinion-oriented answers for a given set of questions. (Yu and Hatzivassiloglou, 2003) separated opinions from facts and summarized them as answer to opinion questions. (Kim and Hovy, 2005) identified opinion holders, which are a key component in retrieving the correct answers to opinion questions. Due to the realized importance of blog data, recent years have also marked the beginning of NLP research focused on the development of opinion QA systems and the organization of international conferences encouraging the creation of effective QA systems both for fact and subjective texts. The TAC 2008[2] QA track proposed a collection

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

of factoid and opinion queries called “rigid list” (factoid) and “squishy list” (opinion) respectively, to which the traditional QA systems had to be adapted. Some participating systems treated opinionated questions as “other” and thus they did not employ opinion specific methods. However, systems that performed better in the “squishy list” questions than in the “rigid list” implemented additional components to classify the polarity of the question and of the extracted answer snippet. The Alyssa system (Shen et al, 2007) uses a Support Vector Machines (SVM) classifier trained on the MPQA corpus (Wiebe et al., 2005), English NTCIR3 data and rules based on the subjectivity lexicon (Wilson et al., 2005). (Varma et al., 2008) performed query analysis to detect the polarity of the question using defined rules. Furthermore, they filter opinion from fact retrieved snippets using a classifier based on Naïve Bayes with unigram features, assigning for each sentence a score that is a linear combination between the opinion and the polarity scores. The PolyU (Venjie et al., 2008) system determines the sentiment orientation of the sentence using the Kullback-Leibler divergence measure with the two estimated language models for the positive versus negative categories. The QUANTA (Li et al., 2008) system performs opinion question sentiment analysis by detecting the opinion holder, the object and the polarity of the opinion. It uses a semantic labeler based on PropBank[4] and manually defined patterns. Regarding the sentiment classification, they extract and classify the opinion words. Finally, for the answer retrieval, they score the retrieved snippets depending on the presence of topic and opinion words and only choose as answer the top ranking results. Other related work concerns opinion holder and target detection. NTCIR 7 and 8 organized MOAT (the Multilingual Opinion Analysis Task), in which most participants employed machine learning approaches using syntactic patterns learned on the MPQA corpus (Wiebe et al., 2005). Starting from the abovementioned research, our aim is to take a step forward to present approaches and employ opinion specific methods focused on improving the performance of our OQA. We perform the retrieval at 1 sentence and 3 sentence-level and also determine the expected source (ES) and the expected target (ET) of the questions, which are fundamental to properly retrieve the correct answer. These two elements are selected employing semantic roles (SR). The expected answer type (EAT) is determined using Machine Learning (ML) using Support Vector Machine (SVM), by taking into account the interrogation formula, the subjectivity of the verb and the presence of polarity words in the target SR. In the case of expected opinionated answers, we also compute the expected polarity type (EPT) – by applying opinion mining (OM) on the affirmative version of the question (e.g. for the question *“Why do people prefer Starbucks to Dunkin Donuts?”*, the affirmative version is *“People prefer Starbucks to Dunkin Donuts because X”*). These experiments are presented in more detail in Section 5.

4 Corpora

In order to carry out the present research for detecting and solving the complexities of opinion QA, we employed two corpora of blog posts: *EmotiBlog* (Boldrini *et al.*, 2009a) and the TAC 2008 Opinion Pilot test collection (part of the Blog06 corpus).

The TAC 2008 Opinion Pilot test collection is composed by documents with the answers to the opinion questions given on 25 targets. *EmotiBlog* is a collection of blog posts in English extracted form the Web. As a consequence, it represents a genuine example of this textual genre. It consists in a monothematic corpus about the Kyoto Protocol, annotated with the improved version of *EmotiBlog* (Boldrini *et al.*, 2009b). It is well know that Opinion Mining (OM) is a very complex task due to the high variability of the language employed. Thus, our objective is to build an annotation model that is able to capture the whole range of phenomena specific to subjectivity expression. Additional criteria employed when choosing the elements to be annotated were effectiveness and noise minimization. Thus, from the first version of the model, the elements which did not prove to be statistically relevant have been eliminated. The elements that compose the improved version of the annotation model are presented in Table 1.

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

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

Elements	Description
Obj. speech	Confidence, comment, source, target.
Subj. speech	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Adjectives/Adverbs	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
Verbs/ Names	Confidence, comment, level, emotion, phenomenon, polarity, mode, source and target.
Anaphora	Confidence, comment, type, source and target.
Capital letter/ Punctuation	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Phenomenon	Confidence, comment, type, collocation, saying, slang, title, and rhetoric.
Reader/Author Interpr. (obj.)	Confidence, comment, level, emotion, phenomenon, polarity, source and target.
Emotions	Confidence, comment, accept, anger, anticipation, anxiety, appreciation, bad, bewilderment, comfort, compassion…

Table 1: *EmotiBlog* improved structure

The first distinction consists in separating objective and subjective speech. Subsequently, a finer-grained annotation is employed for each of the two types of data. Objective sentences are annotated with source and target (when necessary, also the level of confidence of the annotator and a comment). Subjective elements can be annotated at a sentence level, but they also have to be labeled at a word and/or phrase level. *EmotiBlog* also contains annotations of anaphora at a cross-document level (to interpret the storyline of the posts) and the sentence type (simple sentence or title, but also saying or collocation). Finally, the Reader and the Writer interpretation have to be marked in objective sentences. These elements are employed to mark and interpret correctly an apparent objective discourse, whose aim is to implicitly express an opinion (e.g. *"The camera broke in two days"*). The first is useful to extract what is the interpretation of the reader (for example if the writer says *The result of their governing was an increase of 3.4% in the unemployment rate* instead of *The result of their governing was a disaster for the unemployment rate*) and the second to understand the background of the reader (i.e.. *These criminals are not able to govern* instead of saying *the x party is not able to govern*). From this sentence, for example, the reader can deduce the political ideas of the writer. The questions whose answers are annotated with *EmotiBlog* are the subset of opinion questions in English presented in (Balahur et al., 2009). The complete list of questions is shown in Table 2.

N	Question
2	What motivates people's negative opinions on the Kyoto Protocol?
5	What are the reasons for the success of the Kyoto Protocol?
6	What arguments do people bring for their criticism of media as far as the Kyoto Protocol is concerned?
7	Why do people criticize Richard Branson?
11	What negative opinions do people have on Hilary Benn?
12	Why do Americans praise Al Gore's attitude towards the Kyoto protocol?
15	What alternative environmental friendly resources do people suggest to use instead of gas en the future?
16	Is Arnold Schwarzenegger pro or against the reduction of CO2 emissions?
18	What improvements are proposed to the Kyoto Protocol?
19	What is Bush accused of as far as political measures are concerned?
20	What initiative of an international body is thought to be a good continuation for the Kyoto Protocol?

Table 2: Questions over the *EmotiBlog* corpus

The main difference between the two corpora employed is that *Emotiblog* is monothematic, containing only posts about the Kyoto Protocol, while the TAC 2008 corpus contains documents on a multitude of subjects. Therefore, different techniques must be adjusted in order to treat each of them.

5 Experiments

5.1 Question Analysis

In order to be able to extract the correct answer to opinion questions, different elements must be considered. As stated in (Balahur et al., 2009) we need to determine both the expected answer type (EAT) of the question – as in the case of factoid ones - as well as new elements – such as expected polarity type (EPT). However, opinions are directional – i.e., they suppose the existence of a source and a target to which they are addressed. Thus, we introduce two new elements in the question analysis – expected source (ES) and expected target (ET). These two elements are selected by applying SR and choosing the source as the agent in the sentence and the direct object (patient) as the target of the opinion. Of course, the source and target of the opinions expressed can also be found in other roles, but at this stage we only consider these cases. The expected answer type (EAT) (e.g. opinion or other) is determined using Machine Learning (ML) using Support Vector Machine (SVM), by taking into account the interrogation formula, the subjectivity of the verb and the presence of polarity words in the target SR. In the case of expected opinionated answers, we also compute the expected polarity type (EPT) – by applying OM on the affirmative version of the question. An example of such a transformation is: given the question *"What are the reasons for the success of the Kyoto Protocol?"*, the affirmative version of the question is *"The reasons for the success of the Kyoto Protocol are X"*.

5.2 Candidate Snippet Retrieval

In the answer retrieval stage, we employ four strategies:

1. Using the JIRS (JAVA Information Retrieval System) IR engine (Gómez et al., 2007) to find relevant snippets. JIRS retrieves passages (of the desired length), based on searching the question structures (n-grams) instead of the keywords, and comparing them.

2. Using the "Yahoo" search engine to retrieve the first 20 documents that are most related to the query. Subsequently, we apply LSA on the retrieved documents and extract the words that are most related to the topic. Finally, we expand the query using words that are very similar to the topic and retrieve snippets that contain at least one of them and the ET.
3. Generating equivalent expressions for the query, using the DIRT paraphrase collection (Lin and Pantel, 2001) and retrieving candidate snippets of length 1 and 3 (length refers to the number of sentences retrieved) that are similar to each of the new generated queries and contain the ET. Similarity is computed using the cosine measure. Examples of alternative queries for *"People like George Clooney"* are *"People adore George Clooney"*, *"People enjoy George Clooney"*, *"People prefer George Clooney"*.
4. Enriching the equivalent expressions for the query in 3. with the topic-related words discovered in 2. using LSA.

5.3 Polarity and topic-polarity classification of snippets

In order to determine the correct answers from the collection of retrieved snippets, we must filter for the next processing stage only the candidates that have the same polarity as the question EPT. For polarity detection, we use a combined system employing SVM ML on unigram and bigram features trained on the NTCIR MOAT 7 data and an unsupervised lexicon-based system. In order to compute the features for each of the unigrams and bigrams, we compute the tf-idf scores.

The unsupervised system uses the Opinion Finder lexicon to filter out subjective sentences – that contain more than two subjective words or a subjective word and a valence shifter (obtained from the General Inquirer resource). Subsequently, it accounts for the presence of opinionated words from four different lexicons – MicroWordNet (Cerini et al., 2007), WordNet Affect (Strapparava and Valitutti, 2004) Emotion Triggers (Balahur and Montoyo, 2008) and General Inquirer (Stone et al., 1966). For the joint topic-polarity analysis, we first employ LSA to determine the words that are strongly associated to the topic, as described in Section 5.2 (second list item). Consequently, we compute the polarity of the sentences that contain at least one topic word and the question target.

5.4 Filtering using SR

Finally, answers are filtered using the *Semrol* system for SR labeling described in (Moreda, 2008). Subsequently, we filter all snippets with the required target and source as agent or patient. *Semrol* receives as input plain text with information about grammar, syntax, word senses, Named Entities and constituents of each verb. The system output is the given text, in which the semantic roles information of each constituent is marked. Ambiguity is resolved depending on the machine algorithm employed, which in this case is TIMBL[5].

6 Evaluation and Discussion

We evaluate our approaches on both the *EmotiBlog* question collection, as well as on the TAC 2008 Opinion Pilot test set. We compare them against the performance of the system evaluated in (Balahur et al., 2009) and the best (Copeck et al., 2008) and worst (Varma et al., 2008) scoring systems (as far as F-measure is concerned) in the TAC 2008 task. For both the TAC 2008 and *EmotiBlog* sets of questions, we employ the SR system in SA and determine the ES, ET and EPT. Subsequently, for each of the two corpora, we retrieve 1-phrase and 3-phrase snippets. The retrieval of the of the *EmotiBlog* candidate snippets is done using query expansion with LSA and filtering according to the ET. Further on, we apply sentiment analysis (SA) using the approach described in Section 5.3 and select only the snippets whose polarity is the same as the determined question EPT. The results are presented in Table 3.

Q No.	**No.** A	**Baseline** (Balahur et al., 2009)				**1 phrase + ET+SA**				**3 phrases +ET+SA**			
		@1	@5	@10	@50	@1	@5	@10	@50	@1	@5	@10	@20
2	5	0	2	3	4	1	2	3	4	1	2	3	4
5	11	0	0	0	0	0	2	2	2	1	2	3	4
6	2	0	0	1	2	1	1	2	2	0	1	2	2
7	5	0	0	1	3	1	1	1	3	0	2	2	4
11	2	1	1	1	1	0	0	0	0	0	0	0	1
12	3	0	1	1	1	0	1	2	3	0	0	1	2
15	1	0	0	1	1	0	0	1	1	1	1	1	1
16	6	1	4	4	4	0	1	1	2	1	2	2	6
18	1	0	0	0	0	0	0	0	0	0	0	0	0
19	27	1	5	6	18	0	1	1	2	0	1	1	1
20	4	0	0	0	0	0	0	1	1	0	0	1	2

Table 3: Results for questions over EmotiBlog

The retrieval of the TAC 2008 1-phrase and 3-phrase candidate snippets was done using JIRS and, in a second approach, using the cosine similarity measure between alternative queries generated using paraphrases and candidate snippets. Subsequently, we performed different evaluations, in order to assess the impact of using different resources and tools. Since the TAC 2008 had a limit of the output of 7000 characters, in order to compute a comparable F-measure, at the end of each processing chain, we only considered the snippets for the 1-phrase retrieval and for the 3-phases one until this limit was reached.

1. In the first evaluation, we only apply the sentiment analysis tool and select the snippets that have the same polarity as the question EPT and the ET is found in the snippet. (i.e. *What motivates peoples negative opinions on the Kyoto Protocol? The Kyoto Protocol becomes deterrence to economic development and international cooperation/ Secondly, in terms of administrative aspect, the Kyoto Protocol is difficult to implement.* - same EPT and ET)
   We also detected cases of same polarity but no ET, e.g. *These attempts mean annual expenditures of $700 million in tax credits in order to endorse technologies, $3 billion in developing research and $200 million in settling technology into developing countries* – EPT negative but not same ET.
2. In the second evaluation, we add the result of the LSA process to filter out the snippets from 1., containing the words related to the topic starting from the retrieval performed by Yahoo, which extracts the first 20 documents about the topic.
3. In the third evaluation, we filter the results in 2 by applying the *Semrol* system and setting the condition that the ET and ES are the agent or the patient of the snippet.
4. In the fourth evaluation setting, we replaced the set of snippets retrieved using JIRS with the ones obtained by generating alternative queries using paraphrases (as explained in the third method in section 5.2.). We subsequently filtered these results based on their polarity (so that it corresponds to the EPT) and on the condition that the source and target of the opinion (identified through SRL using Semrol) correspond to the ES and ET.

[5] http://ilk.uvt.nl/downloads/pub/papers/Timbl_6.2_Manual.pdf and http://ilk.uvt.nl/timbl/

5. In the fourth evaluation setting, we replaced the set of snippets retrieved using JIRS with the ones obtained by generating alternative queries using paraphrases, enriched with the topic words determined using LSA. We subsequently filtered these results based on their polarity (so that it corresponds to the EPT) and on the condition that the source and target of the opinion (identified through SRL using Semrol) correspond to the ES and ET.

System	F-measure
Best TAC	0.534
Worst TAC	0.101
JIRS + SA+ET (1 phrase)	0.377
JIRS + SA+ET (3 phrases)	0.431
JIRS + SA+ET+LSA (1 phrase)	0.489
JIRS + SA+ET+LSA (3 phrases)	0.505
JIRS + SA+ET+LSA+SR (1 phrase)	0. 533
JIRS + SA+ET+LSA+SR (3 phrases)	0.571
PAR+SA+ET+SR(1 phrase)	0.345
PAR+SA+ET+SR(2 phrase)	0.386
PAR_LSA+SA+ET+SR (1 phrase)	0.453
PAR_LSA+SA+ET+SR (3 phrases)	0.434

Table 4: Results for the TAC 2008 test set

From the results obtained (Table 3 and Table 4), we can draw the following conclusions. Firstly, the hypothesis that OQA requires the retrieval of longer snippets was confirmed by the improved results, both in the case of *EmotiBlog*, as well as the TAC 2008 corpus. Secondly, opinion questions require the use of joint topic-sentiment analysis. As we can see from the results, the use of topic-related words when computing of the affect influences the results in a positive manner and joint topic-sentiment analysis is especially useful for the cases of questions asked on a monothematic corpus. Thirdly, another conclusion that we can draw is that target and source detection are highly relevant steps at the time of answer filtering, not only helping the more accurate retrieval of answers, but also at placing at the top of the retrieval the relevant results (as more relevant information is contained within these 7000 characters). The use of paraphrases at the retrieval stage was shown to produce a significant drop in results, which we explain by the noise introduced and the fact that more non-relevant answer candidates were introduced among the results. Nonetheless, as we can see from the overall relatively low improvement in the results, much remains to be done in order to appropriately tackle OQA. As seen in the results, there are still questions for which no answer is found (e.g. 18). This is due to the fact that the treatment of such questions requires the use of inference techniques that are presently unavailable (i.e. define terms such as *"improvement"*, possibly as *"X better than Y"*, in which case opinion extraction from comparative sentences should be introduced in the model).

The results obtained when using all the components for the 3-sentence long snippets significantly improve the results obtained by the best system participating in the TAC 2008 Opinion Pilot competition (determined using a paired t-test for statistical significance, with confidence level 5%). Finally, from the analysis of the errors, we could see that even though some tools are in theory useful and should produce higher improvements – such as SR – their performance in reality does not produce drastically higher results. The idea to use paraphrases for query expansion also proved to decrease the system performance. From preliminary results obtained using JavaRap[6] for coreference resolution, we also noticed that the performance of the OQA lowered, although theoretically it should have improved.

7 Conclusions ad Future Work

In this paper, we presented and evaluated different methods and techniques with the objective of improving the task of QA in the context of opinion data. From the evaluations performed using different NLP resources and tools, we concluded that joint topic-sentiment analysis, as well as the target and source identification, are crucial for the correct performance of this task. We have also demonstrated that by retrieving longer answers, the results have improved. We tested, within a simple setting, the impact of using paraphrases in the context of opinion questions and saw that their use lowered the system results. Although such paraphrase col-

[6] http://wing.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.htm

lections include a lot of noise and have been shown to decrease system performance even in the case of factual questions, we believe that other types of paraphrasing methods should be investigated in the context of OQA. We thus showed that opinion QA requires the development of appropriate strategies at the different stages of the task (recognition of subjective questions, detection of subjective content of the questions, source and target identification, retrieval and classification of the candidate answer data). Due to the high level of complexity of subjective language, our future work will be focused on testing higher-performing tools for coreference resolution, other (opinion) paraphrases collections and paraphrasing methods and the employment of external knowledge sources that refine the semantics of queries. We also plan to include other SA methods and extend the semantic roles considered for ET and ES, with the purpose of checking if they improve or not the performance of the QA system.

Acknowledgements

This paper has been partially supported by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educación - Generalitat Valenciana (grant no. PROMETEO/2009/119 and ACOMP/2010/286).

References

Balahur, A. and Montoyo, A. 2008. *Applying a Culture Dependent Emotion Triggers Database for Text Valence and Emotion Classification.* In Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine, Aberdeen, Scotland.

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

Balahur, A., Boldrini, E., Montoyo A. and Martínez-Barco P. 2009. *Opinion and Generic Question Answering Systems: a Performance Analysis.* In Proceedings of ACL. Singapur.

Boldrini, E., Balahur, A., Martínez-Barco, P. and Montoyo. A. 2009a. *EmotiBlog: an Annotation Scheme for Emotion Detection and Analysis in Non-traditional Textual Genre.* In Proceedings of DMIN 2009, Las Vegas. Nevada.

Boldrini, E., Balahur, A., Martínez-Barco, P. and Montoyo. A. 2009b. *EmotiBlog: a fine-grained model for emotion detection in non-traditional textual genre.* In Proceedings of WOMSA 2009. Seville.

Cardie, C., Wiebe, J., Wilson, T. and Litman, D. 2003. *Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering.* AAAI Spring Symposium on New Directions in Question Answering.

Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M. and Gandini, C. 2007. *Micro-WNOp: A gold standard for the evaluation of automatically compiledlexical resources for opinion mining*. In: A.Sanso (ed.): Language resources and linguistic theory: Typology, Second Language Acquisition, English Linguistics. Milano. IT.

Copeck, T., Kazantseva, A., Kennedy, A., Kunadze, A., Inkpen, D. and Szpakowicz, S. 2008. *Update Summary Update*. In Proceedings of the Text Analysis Conference (TAC) 2008.

Cui, H., Mittal, V. and Datar, M. 2006. *Comparative Experiments on Sentiment Classification for Online Product Review.* Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference. Boston, Massachusetts, USA.

Gómez, J.M., Rosso, P. and Sanchis, E. 2007. *JIRS Language-Independent Passage Retrieval System: A Comparative Study.* 5th International Conference on Natural Language Proceeding (ICON 2007).

Kabadjov, M., Balahur, A. And Boldrini, E. 2009. *Sentiment Intensity: Is It a Good Summary Indicator?.* Proceedings of the 4th Language Technology Conference LTC, pp. 380-384. Poznan, Poland, 6-8.11.2009.

Kim, S. M. and Hovy, E. 2005. *Identifying Opinion Holders for Question Answering in Opinion Texts.* Proceedings of the Workshop on Question Answering in Restricted Domain at the Conference of the American Association of Artificial Intelligence (AAAI-05). Pittsburgh, PA.

Li, F., Zheng, Z.,Yang T., Bu, F., Ge, R., Zhu, X., Zhang, X., and Huang, M. 2008. *THU QUANTA at TAC 2008. QA and RTE track.* In Proceedings of the Text Analysis Conference (TAC).

Lin, D. and Pantel, P. 2001. *Discovery of Inference Rules for Question Answering.* Natural Language Engineering 7(4):343-360.

Moreda. P. 2008. *Los Roles Semánticos en la Tecnología del Lengauje Humano: Anotación y Aplicación.* Doctoral Thesis. University of Alicante.

Pustejovsky, J. and Wiebe, J. 2006. *Introduction to Special Issue on Advances in Question Answering.* Language Resources and Evaluation (2005), (39).

Shen, D., Wiegand, M., Merkel, A., Kazalski, S., Hunsicker, S., Leidner, J. L. and Klakow, D. 2007. *The Alyssa System at TREC QA 2007: Do We Need Blog06?* In Proceedings of the Sixteenth Text Retrieval Conference (TREC 2007), Gaithersburg, MD, USA.

Strapparava, C. and Valitutti, A. 2004. *WordNet-Affect: an affective extension of WordNet*. In Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004), pages 1083 – 1086, Lisbon.

Stoyanov, V., Cardie, C., and Wiebe, J. 2005. *Multiperspective question answering using the opqa corpus*. In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005).

Varma, V., Pingali, P., Katragadda, S., Krishna, R., Ganesh, S., Sarvabhotla, K. Garapati, H., Gopisetty, H., Reddy, K. and Bharadwaj, R. 2008. *IIIT Hyderabad at TAC 2008*. In Proceedings of Text Analysis Conference (TAC).

Wenjie, L., Ouyang, Y., Hu, Y. and Wei, F. 2008. *PolyU at TAC 2008.* In Proceedings of the Text Analysis Conference (TAC).

Wiebe, J., Wilson, T., and Cardie, C. 2005. *Annotating expressions of opinions and emotions in language.* Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.

Wilson, T., J. Wiebe, and Hoffmann, P. 2005. *Recognizing Contextual Polarity in Phrase-level sentiment Analysis.* In Proceedings of the Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/ EMNLP).

Yu, H. and Hatzivassiloglou, V. 2003. *Towards Answering Opinion Questions: Separating Facts from Opinions.* In Proceedings of EMNLP-03.

Wiebe, J., Wilson, T., and Cardie, C. (2005). *Annotating expressions of opinions and emotions in language.* In Language Resources and Evaluation. Vol. 39.