Opinion and Generic Question Answering Systems: a Performance Analysis

Alexandra Balahur [1,2]
[1]DLSI, University of Alicante
Ap. De Correos 99, 03080, Alicante
[2]IPSC, EC Joint Research Centre
Via E. Fermi, 21027, Ispra
abalahur@dlsi.ua.es

Ester Boldrini
DLSI, University of Alicante
Ap. De Correos 99, 03080, Alicante
eboldrini@dlsi.ua.es

Andrés Montoyo
DLSI, University of Alicante
Ap. De Correos 99, 03080, Alicante
montoyo@dlsi.ua.es

Patricio Martínez-Barco
DLSI, University of Alicante
Ap. De Correos 99, 03080, Alicante
patricio@dlsi.ua.es

Abstract

The importance of the new textual genres such as blogs or forum entries is growing in parallel with the evolution of the Social Web. This paper presents two corpora of blog posts in English and in Spanish, annotated according to the *EmotiBlog* annotation scheme. Furthermore, we created 20 factual and opinionated questions for each language and also the *Gold Standard* for their answers in the corpus. The purpose of our work is to study the challenges involved in a mixed fact and opinion question answering setting by comparing the performance of two Question Answering (QA) systems as far as mixed opinion and factual setting is concerned. The first one is open domain, while the second one is opinion-oriented. We evaluate separately the two systems in both languages and propose possible solutions to improve QA systems that have to process mixed questions.

Introduction and motivation

In the last few years, the number of blogs has grown exponentially. Thus, the Web contains more and more subjective texts. A research from the Pew Institute shows that 75.000 blogs are created daily (Pang and Lee, 2008). They approach a great variety of topics (computer science, sociology, political science or economics) and are written by different types of people, thus are a relevant resource for large community behavior analysis. Due to the high volume of data contained in blogs, new Natural Language Processing (NLP) resources, tools and methods are needed in order to manage their language understanding. Our fist contribution consists in carrying out a multilingual research, for English and Spanish. Secondly, many sources are present in blogs, as people introduce quotes from newspaper articles or other information to support their arguments and make references to previous posts in the discussion thread. Thus, when performing a task such as Question Answering (QA), many new aspects have to be taken into consideration. Previous studies in the field (Stoyanov, Cardie and Wiebe, 2005) showed that certain types of queries, which are factual in nature, require the use of Opinion Mining (OM) resources and techniques to retrieve the correct answers. A further contribution this paper brings is the analysis and definition of the criteria for the discrimination among types of factual versus opinionated questions. Previous researchers mainly concentrated on newspaper collections. We formulated and annotated of a set of questions and answers over a multilingual *blog* collection. A further contribution is the evaluation and comparison of two different approaches to QA a fact-oriented one and another designed for opinion QA scenarios.

Related work

Research in building factoid QA systems has a long history. However, it is only recently that studies have started to focus also on the creation and development of QA systems for opinions. Recent years have seen the growth of interest in this field, both by the research performed and the publishing of various studies on the requirements

and peculiarities of opinion QA systems (Stoyanov, Cardie and Wiebe, 2005), (Pustejovsky and Wiebe, 2006), as well as the organization of international conferences that promote the creation of effective QA systems both for general and subjective texts, as, for example, the Text Analysis Conference (TAC)[1]. Last year's TAC 2008 Opinion QA track proposed a mixed setting of factoid ("rigid list") and opinion questions ("squishy list"), to which the traditional systems had to be adapted. The Alyssa system (Shen *et al*., 2007), classified the polarity of the question and of the extracted answer snippet, using a Support Vector Machines classifier trained on the MPQA corpus (Wiebe, Wilson and Cardie, 2005), English NTCIR[2] data and rules based on the subjectivity lexicon (Wilson, Wiebe and Hoffman, 2005). The PolyU (Wenjie *et al.,* 2008) system determines the sentiment orientation with two estimated language models for the positive versus negative categories. The QUANTA (Li, 2008) system detects the opinion holder, the object and the polarity of the opinion using a semantic labeler based on PropBank[3] and some manually defined patterns.

Evaluation

In order to carry out our evaluation, we employed a corpus of blog posts presented in (Boldrini *et al*., 2009). It is a collection of blog entries in English, Spanish and Italian. However, for this research we used the first two languages. We annotated it using *EmotiBlog* (Balahur *et al.,* 2009) and we also created a list of 20 questions for each language. Finally, we produced the *Gold Standard*, by labeling the corpus with the correct answers corresponding to the questions.

1.1 Questions

No	TYPE		QUESTION
1	F	F	What international organization do people criticize for its policy on carbon emissions? *¿Cuál fue uno de los primeros países que se preocupó por el problema medioambiental?*
2	O	F	What motivates people's negative opinions on the Kyoto Protocol? *¿Cuál es el país con mayor responsabilidad de la contaminación mundial según la opinión pública?*
3	F	F	What country do people praise for not signing the Kyoto Protocol? *¿Quién piensa que la reducción de la contaminación se debería apoyar en los consejos de los científicos?*
4	F	F	What is the nation that brings most criticism to the Kyoto Protocol? *¿Qué administración actúa totalmente en contra de la lucha contra el cambio climático?*
5	O	F	What are the reasons for the success of the Kyoto Protocol? *¿Qué personaje importante está a favor de la colaboración del estado en la lucha contra el calentamiento global?*
6	O	F	What arguments do people bring for their criticism of media as far as the Kyoto Protocol is concerned? *¿A qué políticos americanos culpa la gente por la grave situación en la que se encuentra el planeta?*
7	O	F	Why do people criticize Richard Branson? *¿A quién reprocha la gente el fracaso del Protocolo de Kyoto?*
8	F	F	What president is criticized worldwide for his reaction to the Kyoto Protocol? *¿Quién acusa a China por provocar el mayor daño al medio ambiente?*
9	F	O	What American politician is thought to have developed bad environmental policies? *¿Cómo ven los expertos el futuro?*
10	F	O	What American politician has a positive opinion on the Kyoto protocol? *Cómo se considera el atentado del 11 de septiembre?*
11	O	O	What negative opinions do people have on Hilary Benn? *¿Cuál es la opinión sobre EEUU?*
12	O	O	Why do Americans praise Al Gore's attitude towards the Kyoto protocol and other environmental issues? *¿De dónde viene la riqueza de EEUU?*
13	F	O	What country disregards the importance of the Kyoto Protocol? *¿Por qué la guerra es negativa?*
14	F	O	What country is thought to have rejected the Kyoto Protocol due to corruption? *¿Por qué Bush se retiró del Protocolo de Kyoto?*
15	F/ O	O	What alternative environmental friendly resources do people suggest to use instead of gas en the future? *¿Cuál fue la posición de EEUU sobre el Protocolo de Kyoto?*
16	F/ O	O	Is Arnold Schwarzenegger pro or against the reduction of CO2 emissions? *¿Qué piensa Bush sobre el cambio climático?*
17	F	O	What American politician supports the reduction of CO2 emissions? *¿Qué impresión da Bush?*
18	F/ O	O	What improvements are proposed to the Kyoto Protocol? *¿Qué piensa China del calentamiento global?*
19	F/ O	O	What is Bush accused of as far as political measures are concerned? *¿Cuál es la opinión de Rusia sobre el Protocolo de Kyoto?*
20	F/ O	O	What initiative of an international body is thought to be a good continuation for the Kyoto Protocol? *¿Qué cree que es necesario hacer Yvo Boer?*

Table 1: List of question in English and Spanish

As it can be seen in the table above, we created factoid (F) and opinion (O) queries for English and for Spanish; however, there are some that could be defined between factoid and opinion (F/O) and the system can retrieve multiple answers after having selected, for example, the polarity of the sentences in the corpus.

1.2 Performance of the two systems

We evaluated and compared the generic QA system of the University of Alicante (Moreda *et al.,* 2008) and the opinion QA system presented in (Balahur *et al.,* 2008), in which Named Entity Recognition with LingPipe[4] and FreeLing[5] was

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

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

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

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

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

added, in order to boost the scores of answers containing NEs of the question Expected Answer Type (EAT). Table 2 presents the results obtained for English and Table 3 for Spanish. We indicate the id of the question (Q), the question type (T) and the number of answer of the *Gold Standard* (A). We present the number of the retrieved questions by the traditional system (TQA) and by the opinion one (OQA). We take into account the first 1, 5, 10 and 50 answers.

Q	T	A	Number of found answers							
			@1		@5		@10		@ 50	
			TQA	OQA	TQA	OQA	TQA	OQA	TQA	OQA
1	F	5	0	0	0	2	0	3	4	4
2	O	5	0	0	0	1	0	1	0	3
3	F	2	1	1	2	1	2	1	2	1
4	F	10	1	1	2	1	6	2	10	4
5	O	11	0	0	0	0	0	0	0	0
6	O	2	0	0	0	0	0	1	0	2
7	O	5	0	0	0	0	0	1	0	3
8	F	5	1	0	3	1	3	1	5	1
9	F	5	0	1	0	2	0	2	1	3
10	F	2	1	0	1	0	1	1	2	1
11	O	2	0	1	0	1	0	1	0	1
12	O	3	0	0	0	1	0	1	0	1
13	F	1	0	0	0	0	0	0	0	1
14	F	7	1	0	1	1	1	2	1	2
15	F/O	1	0	0	0	0	0	1	0	1
16	F/O	6	0	1	0	4	0	4	0	4
17	F	10	0	1	0	1	4	1	0	2
18	F/O	1	0	0	0	0	0	0	0	0
19	F/O	27	0	1	0	5	0	6	0	18
20	F/O	4	0	0	0	0	0	0	0	0

Table 2: Results for English

Q	T	A	Number of found answers							
			@1		@5		@10		@ 50	
			TQA	OQA	TQA	OQA	TQA	OQA	TQA	OQA
1	F	9	1	0	0	1	1	1	1	3
2	F	13	0	1	2	3	0	6	11	7
3	F	2	0	1	0	2	0	2	2	2
4	F	1	0	0	0	0	0	0	1	0
5	F	3	0	0	0	0	0	0	1	0
6	F	2	0	0	0	1	0	1	2	1
7	F	4	0	0	0	0	1	0	4	0
8	F	1	0	0	0	0	0	0	1	0
9	O	5	0	1	0	2	0	2	0	4
10	O	2	0	0	0	0	0	0	0	0
11	O	5	0	0	0	1	0	2	0	3
12	O	2	0	0	0	1	0	1	0	1
13	O	8	0	1	0	2	0	2	0	4
14	O	25	0	1	0	2	0	4	0	8
15	O	36	0	1	0	2	0	6	0	15
16	O	23	0	0	0	0	0	0	0	0
17	O	50	0	1	0	5	0	6	0	10
18	O	10	0	1	0	1	0	2	0	2
19	O	4	0	1	0	1	0	1	0	1
20	O	4	0	1	0	1	0	1	0	1

Table 3: Results for Spanish

1.3 Results and discussion

There are many problems involved when trying to perform mixed fact and opinion QA. The first can be the ambiguity of the questions e.g. *¿De dónde viene la riqueza de EEUU?*. The answer can be explicitly stated in one of the blog sentences, or a system might have to infer them from assumptions made by the bloggers and their comments. Moreover, most of the opinion questions have longer answers, not just a phrase snippet, but up to 2 or 3 sentences. As we can observe in Table 2, the questions for which the TQA system performed better were the pure factual ones (1, 3, 4, 8, 10 and 14), although in some cases (question number 14) the OQA system retrieved more correct answers. At the same time, opinion queries, although revolving around NEs, were not answered by the traditional QA system, but were satisfactorily answered by the opinion QA system (2, 5, 6, 7, 11, 12). Questions 18 and 20 were not correctly answered by any of the two systems. We believe the reason is that question 18 was ambiguous as far as polarity of the opinions expressed in the answer snippets ("improvement" does not translate to either "positive" or "negative") and question 20 referred to the title of a project proposal that was not annotated by any of the tools used. Thus, as part of the future work in our OQA system, we must add a component for the identification of quotes and titles, as well as explore a wider range of polarity/opinion scales. Furthermore, questions 15, 16, 18, 19 and 20 contain both factual as well as opinion aspects and the OQA system performed better than the TQA, although in some cases, answers were lost due to the artificial boosting of the queries containing NEs of the EAT (Expected Answer Type). Therefore, it is obvious that an extra method for answer ranking should be used, as Answer Validation techniques using Textual Entailment. In Table 3, the OQA missed some of the answers due to erroneous sentence splitting, either separating text into two sentences where it was not the case or concatenating two consecutive sentences; thus missing out on one of two consecutively annotated answers. Examples are questions number 16 and 17, where many blog entries enumerated the different arguments in consecutive sentences. Another source of problems was the fact that we gave a high weight to the presence of the NE of the sought type within the retrieved snippet and in some cases the name was misspelled in the blog entries, whereas in other NER performed by

FreeLing either attributed the wrong category to an entity, failed to annotate it or wrongfully annotated words as being NEs. Not of less importance is the question duality aspect in question 17. Bush is commented in more than 600 sentences; therefore, when polarity is not specified, it is difficult to correctly rank the answers. Finally, also the problems of temporal expressions and the coreference need to be taken into account.

Conclusions and future work

In this article, we created a collection of both factual and opinion queries in Spanish and English. We labeled the Gold Standard of the answers in the corpora and subsequently we employed two QA systems, one open domain, one for opinion questions. Our main objective was to compare the performances of these two systems and analyze their errors, proposing solutions to creating an effective QA system for both factoid an opinionated queries. We saw that, even using specialized resources, the task of QA is still challenging. Opinion QA can benefit from a snippet retrieval at a paragraph level, since in many cases the answers were not simple parts of sentences, but consisted in two or more consecutive sentences. On the other hand, we have seen cases in which each of three different consecutive sentences was a separate answer to a question. Our future work contemplates the study of the impact anaphora resolution and temporality on opinion QA, as well as the possibility to use Answer Validation techniques for answer re-ranking.

Acknowledgments

The authors would like to thank Paloma Moreda, Hector Llorens, Estela Saquete and Manuel Palomar for evaluating the questions on their QA system. This research has been partially funded by the Spanish Government under the project TEXT-MESS (TIN 2006-15265-C06-01), by the European project QALL-ME (FP6 IST 033860) and by the University of Alicante, through its doctoral scholarship.

References

Alexandra Balahur, Ester Boldrini, Andrés Montoyo, and Patricio Martínez-Barco, 2009. *Cross-topic Opinion Mining for Real-time Human-Computer Interaction.* In Proceedings of the 6th Workshop in Natural Language Processing and Cognitive Science, ICEIS 2009 Conference, Milan, Italy.

Alexandra Balahur, Elena Lloret, Oscar Ferrandez, Andrés Montoyo, Manuel Palomar, Rafael Muñoz. 2008. *The DLSIUAES Team's Participation in the TAC 2008 Tracks.* In Proceedings of the Text Analysis Conference (TAC 2008).

Ester Boldrini, Alexandra Balahur, Patricio Martínez-Barco, and Andrés Montoyo. 2009. *EmotiBlog: An Annotation Scheme for Emotion Detection and Analysis in Non-Traditional Textual Genres.* To appear in Proceedings of the 5th Conference on data Mining. Las Vegas, Nevada, USA.

W. Li, Y. Ouyang, Y. Hu, F. Wei. *PolyU at TAC 2008.* In Proceedings of Human Language Technologies Conference/Conference on Empirical methods in Natural Language Processing (HLT/EMNLP), Vancouver, BC, Canada, 2008.

Fangtao Li, Zhicheng Zheng, Tang Yang, Fan Bu, Rong Ge, Xiaoyan Zhu, Xian Zhang, and Minlie Huang. *THU QUANTA at TAC 2008 QA and RTE track.* In Proceedings of Human Language Technologies Conference/Conference on Empirical methods in Natural Language Processing (HLT/EMNLP), Vancouver, BC, Canada, 2008.

Bo Pang, and Lilian. Lee, *Opinion mining and sentiment analysis.* Foundations and Trends R. In Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–135, 2008.

James Pustejovsky and Janyce. Wiebe. *Introduction to Special Issue on Advances in Question Answering.* In Language Resources and Evaluation (2005) 39: 119–122. Springer, 2006.

Dan Shen, Jochen L. Leidner, Andreas Merkel, Dietrich Klakow. *The Alyssa system at TREC QA 2007: Do we need Blog06?* In Proceedings of The Sixteenth Text Retrieval Conference (TREC 2007), Gaithersburg, MD, USA, 2007

Vaselin, Stoyanov, Claire Cardie, Janyce Wiebe. *Multi-Perspective Question Answering Using the OpQA Corpus.* In Proceedings of HLT/EMNLP. 2005.

Paloma Moreda, Hector Llorens, Estela Saquete, Manuel Palomar. 2008. Automatic Generalization of a QA Answer Extraction Module Based on Semantic Roles. *In: AAI - IBERAMIA,* Lisbon, Portugal, pages 233-242, Springer.

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

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. *Recognising Contextual Polarity in Phrase-level sentiment Analysis.* In Proceedings of Human language Technologies Conference/Conference on Empirical methods in Natural Language Processing (HLT/EMNLP), Vancouver, BC, Canada, 2005.