A Comparative Study of Open Domain and Opinion Question Answering Systems for Factual and Opinionated Queries

Alexandra Balahur, Ester Boldrini, Andrés Montoyo, Patricio Martínez-Barco
Natural Language Processing and Information Systems Group, Department of Software and Computing Systems.
Apartado de Correos 99, 03080, Alicante

{abalahur, eboldrini, montoyo, patricio}@dlsi.ua.es

Abstract

The development of the Web 2.0 led to the birth of new textual genres such as blogs, reviews or forum entries. The increasing number of such texts and the highly diverse topics they discuss make blogs a rich source for analysis. This paper presents a comparative study on open domain and opinion QA systems. A collection of opinion and mixed fact-opinion questions in English is defined and two Question Answering systems are employed to retrieve the answers to these queries. The first one is generic, while the second is specific for emotions. We comparatively evaluate and analyze the systems' results, concluding that opinion Question Answering requires the use of specific resources and methods.

Keywords Question Answering, Multi-perspective Question Answering, Opinion Annotation, Opinion Mining, Non-Traditional Textual Genres.

1. Introduction

Recent years' statistics show that the number of blogs has been increasing at an exponential rate. A research of the Pew Institute [1] shows that 2-7% of Internet users created a blog and that 11% usually read them. Moreover, researches in different fields proved that this new textual genre is a valuable resource for large community behavior analysis, since blogs address a great variety of topics from a high diversity of social spheres. A common belief is that they are written in a colloquial style, but [2] shows that the language of these texts is not restricted to the more informal levels of expression and a large number of different genres are involved. As a consequence, free expressions, literary prose and newspaper writing coexist without a clear predominance. When using this textual genre, people tend to express themselves freely, using colloquial expressions employed only in day-by-day conversations. Moreover, they can introduce quotes from newspaper articles, news or other sources of information to support their arguments, make references to previous posts or the opinion expressed by others in the discussion thread. Users intervening in debates over one specific topic are from different geographical regions and belong to diverse cultures. All the abovementioned features make blogs a valuable source of information that can be exploited for different purposes. However, due to their language being heterogeneous, it is complex to understand and formalize in order to create effective Natural Language Processing (NLP) tools. At the same time, due to the high volume of data contained in blogs, automatic NLP systems are needed to manage the language understanding and generation. Analyzing emotions and/ or opinions expressed in blog posts could also be useful to predict people's opinion or preferences about a product or an event. One of the other possible applications is an effective Question Answering (QA) system, able to recognize different queries and give the correct answer to both factoid and opinion questions.

2. Related work

QA is the task in which, given a set of questions and a collection of documents where the answers can be found, an automatic NLP system is employed to retrieve the answer to these queries in Natural Language. The main difference between QA and Information Retrieval (IR) is that in the first one, the system is supposed to output the exact answer snippet, whereas in the second task whole paragraphs or even documents are retrieved. Research in building factoid QA systems has a long tradition; however, it is only recently that studies have started to focus on the creation and development of opinion QA systems. Recent years have seen the growth of interest in this field, both by the research and publishing of studies on the requirements and peculiarities of opinion QA systems [4] as well as the organization of international conferences that promote the creation of effective QA systems both for general and subjective texts, such as the Text Analysis Conference (TAC)[1]. Last year's TAC 2008 Opinion QA track proposed a mixed setting of factoid and opinion questions (so called "rigid list" and "squishy list"), to which the traditional systems had to be adapted. Participating systems employed different resources, techniques and methods to overcome the newly introduced difficulties related to opinion mining and polarity classification. The Alyssa system [5], which performed better in the "squishy list" questions than in the

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

International Conference RANLP 2009 - Borovets, Bulgaria, pages 18–22

"rigid list" questions, had additional components implemented for classifying the polarity of the question and of the extracted answer snippet, using a Support Vector Machines (SVM) classifier trained on the MPQA corpus [6], English NTCIR[2] data and rules based on the subjectivity lexicon [7]. Another system introducing new modules to tackle opinion is [8]. They perform query analysis to detect the polarity of the question using defined rules. 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 [9] system determines the sentiment orientation of the sentence and it uses the Kullback-Leibler divergence measure with the two estimated language models for the positive versus negative categories. The UOFL system [10] generates a non-redundant summary of the query for the opinion questions, to take into consideration all the information present in the question, and not only the separated words.

3. Motivation and contribution

Opinion Mining is the task of extracting, given a collection of texts, the opinion expressed on a given target within the documents. It has been proven that performing this task, several other subtasks of NLP can be improved: *Information Extraction* (where opinion mining techniques can be used as a preprocessing step to separate among factual and subjective information), *Authorship Determination* (as subjective language can be considered as a personality mark), *Word Sense Disambiguation*, *multi-source (multi-perspective) summarization* and more informative Answer Retrieval for definition questions [16] (as it can constitute a measure for *credibility*, *sentiment* and *contradictions).* Related work presented research in determining the differences in the characteristics of the fact versus opinion queries and their corresponding answers [11]. However, certain types of questions, which are factual in nature, require the use of Opinion Mining resources and techniques in order to retrieve the correct answers. Our first contribution relies in the analysis and definition of the criteria for the discrimination among different types of factual versus opinionated questions. Furthermore, we created and annotated a set of questions and answers over a multilingual blog collection for English and Spanish. Thus, we also analyze the effect of the textual genre characteristics on the properties of the opinion answers retrieved/missed. A further contribution lies in the evaluation of two different approaches to QA; one is fact oriented (based on Named Entities –NEs–) and the other is specifically designed for opinion QA scenarios. We analyze their different elements, specifications, behavior, evaluated performance and present conclusions on the needs and requirements of systems designed for the presented categories of questions. Last, but not least, using the annotated answers and their corresponding corpus, we analyze possible methods for keyword expansion in an opinion versus fact setting. We present some possible solutions to the shortcomings of direct keyword expansion for opinion QA, employing "polarity-based" expansion using our corpus annotations.

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

4. Corpus collection and analysis

The corpus we employed for our evaluation is composed of blog posts extracted form the Web. It has been collected taking into account the requirements of coherence, authenticity, equilibrium and quality. Our main purpose was to collect a corpus in which the blog posts were about a topic, forming a coherent discussion. Moreover, our collection had to provide a real example of this textual genre, it had to be of the same length for each topic and language, and originated from reliable Web sites. We selected three topics: the Kyoto Protocol, the 2008 Zimbabwe and the USA elections. After having collected the three corpora, we analyzed the characteristic of this textual genre also looking for the subjective expressions and for the way they are formulated in NL. The following step of our research consisted in building up the initial version of *EmotiBlog* [18], an annotation scheme focused on emotions detection in non-traditional textual genres. The annotation scheme is briefly presented in the following section.

5. Annotation scheme

As we mentioned in the previous section, *EmotiBlog* [12] is an annotation scheme for detecting opinion in non-traditional textual genres. It is the first version of a fine-grained and multilingual annotation model that could be useful for an exhaustive comprehension of NL. The first version has been created for English, Italian and Spanish; however, it could be easily adapted for the annotation of other languages. Firstly, we detect the overall sentiment of the blogs and subsequently a distinction between objective and subjective sentences is done. Moreover, for each element, we annotate the source, the target and also a wide range of attributes for the elements (sentiment type, its intensity and polarity, for example). Sentiments are grouped according to [13], who created an alternative dimensional structure of the semantic space for emotions grouping emotions between obstructive and conductive, and finally, between high power and low power control. The annotation task has been carried out by two non-native speakers with extensive knowledge of Spanish and English. The labeling of the 100 texts took approximately one month and a half, working in a part-time schedule. Finally, the last step consisted in labeling the answers to our list of questions to

create a *gold standard* for detecting the mistakes of the QA systems presented in the next section. The list of questions is composed by 20 factual and opinionated queries. Table 1 shows the list of questions.

Table 1: Example of questions

NUM	TYPE	QUESTION
1	F	What international organization do people criticize for its policy on carbon emissions?
2	O	What motivates people's negative opinions on the Kyoto Protocol?
3	F	What country do people praise for not signing the Kyoto Protocol?
4	F	What is the nation that brings most criticism to the Kyoto Protocol?
5	O	What are the reasons for the success of the Kyoto Protocol?
6	O	What arguments do people bring for their criticism of media as far as the Kyoto Protocol is concerned?
7	O	Why do people criticize Richard Branson?
8	F	What president is criticized worldwide for his reaction to the Kyoto Protocol?

As we can see in Table 1, we have a list of opinionated and factoid queries. Factual need a name, date, time, etc as answer, while opinionated ones something more complex. The system should be able firstly to recognize the subjective expressions and after that, discriminate them in order to retrieve the correct answer. In this case the answer can be expressed by an idiom, a saying, or by a sentence and as a consequence it is not a simple name or a date. It is complex because it could be everything; there are no fixed categories of answer types for opinionated questions. As a consequence, we formulated the opinion questions explicitly in order not to increase the difficulty level of the analysis.

6. Evaluation

6.1 Open QA system

With the purpose of evaluating the performance of a general QA system in a mixed fact and opinion setting, we used the QA system of the University of Alicante [14] [15]. It is an open domain QA system employed to deal with factual questions both for English and Spanish. The queries this system can support are *location*, *person*, *organization*, *date-time* and *number*. Furthermore, its architecture is divided into three modules. The first one is the Question Analysis in which the language object of the study is determined using dictionaries with the criterion of selecting the language for which more words are found. Therefore, the question type is selected using a set of regular expressions and the keywords of each question are obtained with morphological and dependencies analysis. For that purpose, MINIPAR[3] for Spanish and Freeling[4] for English are used. The second module is the IR in which the system, originally, relied on the Internet search engines. However, in order to look for information among the Web Log collection, an alternative approach has been developed. A simple keyword-based document retrieval method has been implemented in order to get relevant documents given the question keywords. The last module is called Answer Extraction (AE). The potential answers are selected using a NE recognizer for each retrieved document. LingPipe[5] and Freeling have been used for English and Spanish respectively. Furthermore, NE of the obtained question type and question keywords are marked up in the text. Once selected they are scored and ranked using answer-keywords distances approach. Finally, when all relevant documents have been explored, the system carries out an answer clustering process which groups all answers that are equal or contained by others to the most scored.

6.2 Specific QA system

For the opinion specific QA system, our approach was similar to [16]. Given an opinion question, we try to determine its *polarity*, the *focus*, its *keywords* (by eliminating stopwords) and the *expected answer type* (EAT) (while also marking the NE appearing in it); once this information is extracted from the question, blog texts are split into sentences and NE are marked. Finally, sentences in the blogs are sought which have the highest similarity score with the question keywords, whose polarity is the same as the determined question polarity and which contains a NE of the EAT. As the traditional QA system outputs 50 answers, we also take the 50 most similar sentences and extract the NEs they contain. In the future, when training examples will be available, we plan to set a threshold for similarity, thus not limiting the number of output answers, but setting a border to the similarity score (this is related to the observation in [4] that opinion questions have a highly variable number of answers. In order to extract the topic and determine the question polarity, we define question patterns. These patterns take into consideration the interrogation formula and extract the opinion words (nouns, verbs, adverbs, adjectives and their determiners). They are then classified to determine the polarity of the question, using the WordNet Affect emotion lists, the emotion triggers resource [17], a list of four attitudes containing the verbs, nouns, adjectives and adverbs for the categories of **criticism**, **support**, **admiration** and **rejection** and a list of positive and negative opinion words taken from the system in [18]. On the other hand, we preprocessed the blog texts in order to prepare the answer retrieval. Starting from the focus, keywords and topic of the question, we sought sentences in

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

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

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

the blog collection (which was split into sentences and where Named Entity Recognition was performed using LingPipe) that could constitute possible answers to the questions, according to their similarity to the latter. The similarity score was computed with Pedersen's Text Similarity Package[6]. The condition we subsequently set was that the polarity of the retrieved snipped be the same as the one of the question and, in the case of questions with EAT PERSON, ORGANIZATION or LOCATION, that a Named Entity of the appropriate type was present in the retrieved snippets. In case retrieved snippets containing Named Entities in the question were found, their score was boosted to the score of the most similar snippet retrieved. In case more than 50 snippets were retrieved, we only considered for evaluation the first 50 in the order of their polarity score (which proved to be a good indicator of the snippet's importance [22].

6.3 Evaluation process

We evaluate the performance of the two QA systems in terms of the number of found answers within the top 1, 5, 10 and 50 output answers (TQA is the indicator for the traditional QA system and OQA is the indicator for the opinion QA system). In Table 2 we present the results of the evaluations in the case of each of the 20 questions (the table also contains the type of each questions – F (factual) and O (opinion)). The first observation we can make is the fact that the traditional QA system was able to answer only 8 of the 20 questions we formulated. We will thus compare the performance of the systems at the level of these 8 questions they both answered and separately analyze the faults and strong points, as well as the difficulties of each individual question separately).

Table 2: The QA systems' performance

Question	Type	Number of answers	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

As we can observe in Table 2, as expected, the questions for which the traditional QA system performed better were the pure factual ones (1, 3, 4, 8, 10 and 14), although in some cases (like the one of question number 14) the OQA system retrieved more correct answers. At the same time, purely opinion questions, 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), taking into consideration that a purely word-overlap approach was taken. Questions 18 and 20 were not correctly answered by any of the two systems. We believe this is due to the fact 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. 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. Therefore, it is obvious that an extra method for answer ranking should be used, as Answer Validation techniques using Textual Entailment.

7. Issues and discussion

There are many problems involved when trying to perform opinion QA. Explanations for this fact include ambiguity of the questions (*What is the nation that brings most criticism to the Kyoto Protocol?* – the answer can be explicitly stated in one of the blog sentences, or a system might have to infer them; therefore, the answer is highly contextual and depends on the texts one is analyzing, the need for extra knowledge on the NEs (i.e. *Al Gore is an American politician – should we first look for people that are in favor of environmental measures and test which one is an American politician?)* and the fact that, as opposed to purely factoid questions, most of the opinion questions have answers longer than a single sentence. In many of the cases, the opinion mining system missed on the answers due to erroneous sentence splitting. 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 NER performed by LingPipe either attributed the wrong category to an entity, failed to annotate

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

it or wrongfully annotated words as being NEs when that was not the case. As we could notice, problems of temporal expressions and the coreference need to be taken into account in order to retrieve the correct answer. In most of the time, the QA system need to understand the temporal context of the questions and also of the sentences that compose the corpus, because the present President the USA is different from two years ago, for example. At the other hand, an effective coreference resolution system is indispensable to understand some retrieved answers.

8. Conclusions and future work

In this article, we first presented *EmotiBlog*, an annotation scheme for opinion annotation in blogs and the blog posts collection we gathered to label with our scheme. Subsequently, we presented the collection of mixed opinion and fact questions we created, whose answers we annotated in our corpus. We finally evaluated and discussed on the results of two different QA systems, one that is fact oriented and one that is designed for opinion question answering. Some conclusions that we draw from this analysis are that, even when using specialized resources, the task of opinion QA is still difficult and extra techniques and methods have to be investigated in order to solve the problems we found, parallel to a deeper analysis of the issues involved in this type of QA. In many cases, opinion QA can benefit from a snippet retrieval at a paragraph level, since usually the answers were not mere parts of sentences, but consisted in two or more consecutive sentences. On the other hand, however, we have seen cases in which each of three different consecutive sentences was a separate answer to a question. Future work includes the study of the impact anaphora resolution has on the task of opinion QA, as well as the possibility to use Answer Validation techniques in order to increase the system's performance by answer re-ranking.

9. 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.

10. References

[1] A. Lenhart, J. Horrigan, D. Fallow, *Content Creation Online, Pew Internet & American Life Project*. Available at www.pewinternet.org/pdfs/PIP_Content_Creation_Report.pdf

[2] B. Pang, and L. Lee, *Opinion mining and sentiment analysis. Foundations and Trends R*_In Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–135, 2008.

[3] 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.

[4] V., Stoyanov, C., Cardie, J., Wiebe. *Multi-Perspective Question Answering Using the OpQA Corpus*. HLT/EMNLP. 2005.

[5] D. Shen,. M. Wiegand, A. Merkel, S. Kazalski, S. Hunsicker, J.L. Leidner, and D. 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.

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

[7] T. Wilson, J.Wiebe, and P. 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.

[8] V. Varma, P. Pingali, R. Katragadda, S. Krishna, S. Ganesh, K. Sarvabhotla, H. Garapati, H. Gopisetty, K. Reddy and R. Bharadwaj. In Proceedings of Text Analysis Conference, at the joint annual meeting of TAC and TREC, Gaithersburg, Maryland, USA, 2008.

[9] L. Wenjie, 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.

[10] Y. Chali, S.A. Hasan, S.R. Joty. *(University of Lethbridge) UofL: QA, Summarization (Update Task)*. In Proceedings of Human Language Technologies Conference/Conference on Empirical methods in Natural Language Processing (HLT/EMNLP), Vancouver, BC, Canada, 2008.

[11] V. Stoyanov, C. Cardie, J. Wiebe. *Multi-Perspective Question Answering using the OpQA corpus*. In Proceedings of EMNLP 2005.

[12] A. Balahur, E. Boldrini, A. Montoyo and P. Martínez- Barco. *Cross-topic Opinion Mining for Real-time Human-Computer Interaction*. To appear in Proceedings of ICEIS 2009 Conference, Milan, Italy, 2009.

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

[14] P. Moreda, H. Llorens, E. Saquete, M. Palomar. *The influence of semantic roles in QA: a comparative analysis*. In Proceedings of the SEPLN. MAdris, Spain, pages 55-62, 2008.

[15] P. Moreda, H. Llorens, E. Saquete, M. Palomar. *Automatic Generalization of a QA Answer Extraction Module Based on Semantic Roles.* In: AAI - IBERAMIA, Lisbon, Portugal, pages 233-242, Springer, 2008.

[16] Balahur, A., Lloret, E., Ferrandez, O., Montoyo, A., Palomar, M., Munoz, R. *The DLSIUAES Team's Participation in the TAC 2008 Tracks.* Proceedings of the Text Analysis Conference 2008.

[17] A., Balahur and A., Montoyo. *Applying a culture dependent emotion triggers database for text valence and emotion classification.* In Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 107-114. 2008a.

[18] A., Balahur, A., Montoyo. Multilingual Feature-Driven Opinion Extraction and Summarization from Customer Reviews. In Proceedings of NLDB 2008 – LNCS 5039, pp. 345-346. 2008b.