The OpAL System at NTCIR 8 MOAT

Alexandra Balahur
University of Alicante, DLSI
Ap.de Correos 99
E-03080 Alicante, Spain
0034 965 90 37 72
abalahur@dlsi.ua.es

Ester Boldrini
University of Alicante, DLSI
Ap.de Correos 99
E-03080 Alicante, Spain
0034 965 90 37 72
eboldrini@dlsi.ua.es

Andrés Montoyo
University of Alicante, DLSI
Ap.de Correos 99
E-03080 Alicante, Spain
0034 965 90 37 72
montoyo@dlsi.ua.es

Patricio Martínez-Barco
University of Alicante, DLSI
Ap.de Correos 99
E-03080 Alicante, Spain
0034 965 90 37 72
patricio@dlsi.ua.es

ABSTRACT

The present is marked by the availability of large volumes of heterogeneous data, whose management is extremely complex. While the treatment of factual data has been widely studied, the processing of subjective information still poses important challenges. This is especially true in tasks that combine Opinion Analysis with other challenges, such as the ones related to Question Answering. In this paper, we describe the different approaches we employed in the NTCIR 8 MOAT monolingual English (opinionatedness, relevance, answerness and polarity) and cross-lingual English-Chinese tasks, implemented in our OpAL system. The results obtained when using different settings of the system, as well as the error analysis performed after the competition, offered us some clear insights on the best combination of techniques, that balance between precision and recall. Contrary to our initial intuitions, we have also seen that the inclusion of specialized Natural Language Processing tools dealing with Temporality or Anaphora Resolution lowers the system performance, while the use of topic detection techniques using faceted search with Wikipedia and Latent Semantic Analysis leads to satisfactory system performance, both for the monolingual setting, as well as in a multilingual one.

Categories and Subject Descriptors

[H.3.1 [Information Storage and Retrieval]: linguistic processing

General Terms

Experimentation.

Keywords

Opinion Analysis, Sentiment Analysis, Cross-Lingual Opinion Analysis, Polarity Classification.

1. INTRODUCTION

Together with the growth in the access to technology and the development of the Social Web, the past few years have brought about an interesting and challenging phenomenon – the exponential growth of subjective data on the Web containing both factoid and opinionated information. The fact that anyone can express opinions on anything, on blogs, forums, e-commerce sites, can communicate and share information using social networks, has made data on the Web highly dynamical and growing exponentially. The demonstrated impact such subjective information has on the lives of people everywhere and on the business made the study of automatic processing methods for subjective text an active research field. The Natural Language Processing (NLP) task dealing with the treatment of subjective data is called Sentiment Analysis (SA).

Users must be able to efficiently access this data, through queries. While techniques to retrieve objective information have been widely studied and implemented, opinion-related tasks still represent an important challenge.

2. MOTIVATION AND CONTRIBUTION

It is well known that the task of QA focused on retrieving factoid information has been widely studied, while the treatment of subjective data still remains a challenge. The TAC 2008 Opinion Pilot task[1], as well as the subsequent research performed on the competition data have shown that answering correctly to opinionated questions is different from performing the same tasks in the context of factual data. The first motivation of our work is the urgent need for an effective OQA able to retrieve subjective information that will be employed for a wide range of practical applications. There is the need to detect and explore the challenges raised by Opinion Question Answering (OQA). The first contribution of this paper is a deep study of the performance of retrieval techniques that are not specifically designed for the opinion scenario in the context of these subtasks. Although there is a wide range of tools and methods available implemented for factoid data, the need for specialized tools for opinion information is important and when possible they must be adapted to the needs of the subjective discourse. Another contribution this paper brings is a deep study of the performance of new sentiment-topic detection methods and the introduction of specialized tools for temporal expression and anaphora resolution and the analysis of

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

their effect. Finally, we also introduced new retrieval techniques such as faceted search using Wikipedia with Latent Semantic Analysis (LSA), which demonstrate to improve the performance of the task.

All the above mentioned contributions have been implemented in the OpAL system, with which we participated in the NTCIR 8 MOAT English monolingual and English-Chinese cross-lingual subtasks, obtaining promising results [1].

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 studies have started to focus on the development of OQA systems. Example of this can be [2] who took advantage of opinion summarization to support Multi-Perspective QA system, aiming at extracting opinion-oriented information of a question. [3] separated opinions from facts and summarized them as answer to opinion questions. [4] identified opinion holders, which are frequently asked in 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] Opinion QA track proposed a collection of factoid and opinion queries called “rigid list” (factoid) and “squishy list”(opinion) respectively, to which the traditional 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 [5] uses a Support Vector Machines (SVM) classifier trained on the MPQA corpus [6], English NTCIR[3] data and rules based on the subjectivity lexicon [7]. [8] 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 [9] 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 [10] system performs opinion question sentiment analysis by detecting the opinion holder, the object and the polarity of the opinion. It uses a semantic labeller 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.

4. ENGLISH MONOLINGUAL SUBTASKS

For the English monolingual subtask, the participants were provided with twenty topics. For each of the topics, a question was given, together with a short and concise query, the expected polarity of the answer and the period of time required. For each of the topics, the participants were given a set of documents, that were split into sentences (for the opinionated and relevance judgements) and into opinion units (for the polarity, opinion target and source tasks). We submitted three runs of the OpAL system, for the opinionated, relevance and polarity judgement tasks.

4.1 Judging sentence opinionatedness

The “opinionated” subtask required systems to assign the values YES or NO (Y/N) to each of the sentences in the document collection provided. This value is given depending on whether the sentence contains an opinion (Y) or it does not (N).

In order to judge the opinionatedness of the sentence, we employed two different approaches (the first one corresponding to system run number 1 and the second to system runs 2 and 3).
Both approaches are rule-based, but they differ in the resources employed. We considered as opinionated sentences the ones that contain at least two opinion words or one opinion word preceded by a modifier. For the first approach, the opinion words were taken from the General InquIerer[i], Micro WordNet Opinion and Opinion Finder lexicon and in the second approach we only used the first two resources.

4.2 Determining sentence relevance

In the sentence relevance judgement task, the systems had to out put, for each sentence in the given collection documents per topic, an assessment on whether or not the sentence is relevant for the given question. For the sentence relevance judgement task stage, we employ three strategies (corresponding to the system runs 1,2 and 3, respectively):

1. Using the JIRS (JAVA Information Retrieval System) IR engine [11] 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 faceted search in Wikipedia and performing Latent Semantic Analysis (LSA) to find the words that are most related to the topic. The idea behind this approach is to find the concepts that are contained in the query descriptions of the topics. In order to perform this task, we match the query words, starting from the first, to a category in Wikipedia. Subsequently we match each group of two consecutive words to the same categories, then groups of 3, 4, etc. until the highest match is found. The concepts determined through this process are considered as the topic components. For each of these topic components, we determine the most related words, applying LSA is to the first 20 documents that are retrieved using the Yahoo search engine, given the query. For LSA, we employ the Infomap NLP[5] software. Finally, we expand query using words that are very similar to the topic (retrieved through the LSA process)

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

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

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

[5] http://infomap-nlp.sourceforge.net/

and retrieve snippets that contain at least two such words.

3. The third approach consists in judging, apart from the topic relevance characteristic, the temporal appropriateness of the given sentences. In order to perform this check, we employ TERSEO [12]. We then filter the sentences obtained in the second approach depending on whether or not the document in which they appear have a date matching the required time interval or the sentence with the resolved temporal expressions contains a reference to the required time interval.

4.3 Polarity and topic-polarity classification for judging sentence answerness

The polarity judgment task required the system to assign a value of POS, NEG or NEU (positive, negative or neutral) to each of the sentences in the documents provided. In order to determine the polarity of the sentences, we passed each sentence through an opinion mining system employing SVM machine learning over the NTCIR 7 MOAT corpus, the MPQA corpus and EmotiBlog. Each sentence is preprocessed using Minipar[6]. For the system training, the following features were considered, for each sentence word:

- the part of speech (POS)
- opinionatedness/intensity - if the word is annotated as opinion word, its polarity, i.e. 1 and -1 if the word is positive or negative, respectively and 0 if it is not an opinion word, its intensity (1.2 or 3) and 0 if it is not a subjective word, its emotion (if it has, none otherwise)
- syntactic relatedness with other opinion word – if it is directly dependent of an opinion word or modifier (0 or 1), plus the polarity/intensity and emotion of this word (0 for all the components otherwise).

The difference between the submitted runs consisted in the lexicons used to determine whether a word was opinionated or not. For the first run, we employed the General Inquirer, MicroWordNet and the Opinion Finder opinion resources. For the second one, we employed, aside from these three sources, the "emotion trigger" resource [13].

5. ENGLISH-CHINESE CROSS-LINGUAL SUBTASK

In the Cross-lingual setting, the task of the participating systems was to output, for each of the twenty topics and their corresponding questions (in a language), the list of sentences containing answers (in another language). For this task, we submitted three runs of the OpAL system, all of them for the English- Traditional Chinese cross-lingual setting (i.e. the topics and questions are given in English; the output of the system contains the sentences in set of documents in Traditional Chinese which contain an answer to the given topics).

In the following part, we explain the approaches we followed for each of the system runs.

Given that we had no previous experience with processing Chinese text, the approaches taken were quite simple.

The first step we performed was to tokenize the Chinese texts using LingPipe[7]. Further on, we applied a technique known as "triangulation" to obtain opinion and subjectivity resources for Chinese. The idea behind this approach is to obtain resources for different languages, starting from correct parallel resources in 2 initial languages. The process is exemplified in Figure 1 for obtaining resources in Chinese, starting with resources in English and Spanish.

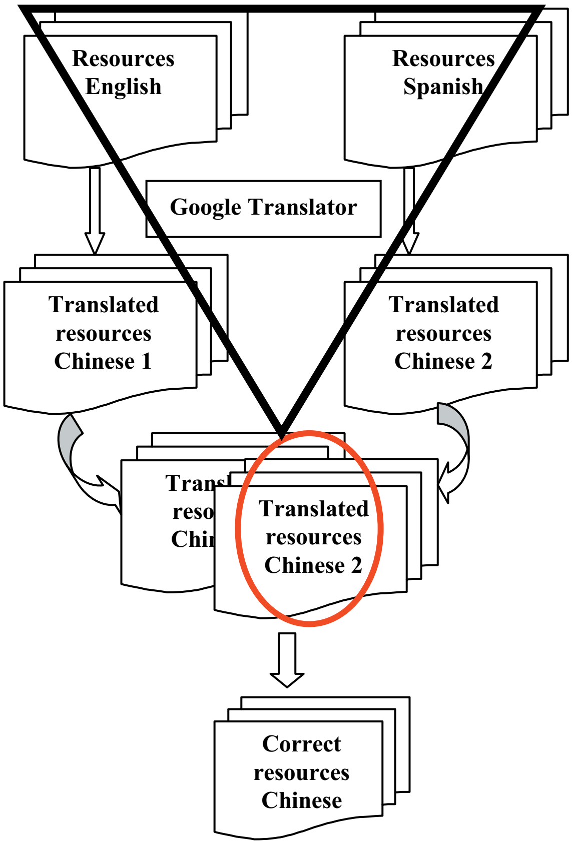
Resources English
Resources Spanish
Google Translator
Translated resources Chinese 1
Translated resources Chinese 2
Transl reso Chi
Translated resources Chinese 2
Correct resources Chinese

Figure 1: Obtaining new resources in Chinese through triangulation

As mentioned before, this technique requires the existence of two correct parallel resources in two different languages to obtain correct resources for a third language. We have previously translated and cleaned the General Inquirer[8], MicroWordNet [14] and Opinion Finder [15] lexicons for Spanish. The "emotion triggers" resource is available both for English, as well as for Spanish. In order to obtain these resources for Traditional Chinese, we use the Google translator. We translate both the English, as well as the Spanish resources, into Traditional Chinese. Subsequently, we performed the intersection of the obtained translations – that is, the corresponding words that have been translated in the same manner – both from English as well as

[6] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm

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

[8] http://www.wjh.harvard.edu/~inquirer/

from Spanish. We removed words that we translated differently from English and Spanish. The intersection words were considered as "clean" (correct) translations. We mapped each of these resources to four classes, depending on the score they are assigned in the original resource – of "high positive", "positive", "high negative" and "negative" and we give each word a corresponding value (4, 1, -4 and -1), respectively.

On the other hand, we translated the topic words determined in English using LSA.

For each of the sentence, we compute a score, given by the sum of the values of the opinion words that are matched in it. In order for a sentence to be considered as answer to the given question, we set the additional conditions that it contains at least one topic word and that the polarity determined corresponds to the required polarity, as given in the topic description. The three runs differ in the resources that were employed to calculate the sentiment score: in the first run, we employed the General Inquirer and MicroWordNet resources; in the second run we added the "emotion trigger resource" and the third run used only the Opinion Finder lexicon.

6. EVALUATION AND DISCUSSION

The following tables present the results of the system runs for the three subtasks in English in which we took part and the cross-lingual English - Traditional Chinese task.

Table 1: Results of system runs for opinionatedness

System RunID	P	R	F
OPAL 1	17.99	45.16	25.73
OPAL 2	19.44	44	26.97
OPAL 3	19.44	44	26.97

Table 2: Results of system runs for relevance

System RunID	P	R	F
OPAL 1	82.05	47.83	60.43
OPAL 2	82.61	5.16	9.71
OPAL 3	76.32	3.94	7.49

Table 3: Results of system runs for polarity

System RunID	P	R	F
OPAL 1	38.13	12.82	19.19
OPAL 2	50.93	12.26	19.76

Table 4: Results of system runs for the cross-lingual task – agreed measures, Traditional Chinese

System RunID	P	R	F
OPAL 1	3.54	56.23	6.34
OPAL 2	3.35	42.75	5.78
OPAL 3	3.42	72.13	6.32

Table 5: Results of system runs for the cross-lingual task – non-agreed measures, Traditional Chinese

System RunID	P	R	F
OPAL 1	14.62	60.47	21.36
OPAL 2	14.64	49.73	19.57
OPAL 3	15.02	77.68	23.55

From the results obtained, we can see that although the extensive filtering according to the topic and the temporal restrictions increases the system precision, we obtain a dramatic drop in the recall. On the other hand, the use of simpler methods in the cross-lingual task yielded better results, the OpAL cross-lingual run 3 obtaining the highest F score for the non-agreed measures and ranking second according to the agreed measures.

From the error analysis performed, we realized that, on the one hand, the LSA-based method to determine topic-related words is not enough to perform this task. The terms obtained by employing this method are correct and useful, but they should be expanded using language models, to better account for the language variability.

Finally, we have seen that systems performing finer tasks, such as temporal expression resolution, are not mature enough to be employed in such tasks. This was confirmed by in-house experiments using anaphora resolution tools such as JavaRAP[9], whose use also led to lower performances of the system and dramatic loss in recall.

7. CONCLUSIONS AND FUTURE WORK

In this paper our research was focused on solving a recent problem born with the massive usage of the Web 2.0: the exponential growth of the subjective data that need to be efficiently managed for a wide range of practical applications. We identified and explored the challenges raised by OQA, as opposed to the traditional QA. Moreover, we studied the performance of new sentiment-topic detection methods and analysed the improvements that can be brought at the different stages of the OQA process and analysed the contribution of discourse analysis, employing techniques such as coreference resolution and temporality detection. We also experimented new retrieval techniques such as faceted search using Wikipedia with LSA, which demonstrate to improve the performance of the task. From the results obtained, we can draw the following conclusions. The first one is that on the one hand, the extensive filtering according to the topic and the temporal restrictions increases the system precision but it produces a dramatic drop in the recall. As a consequence, the use of simpler methods in the cross-lingual task would be more appropriate in this context. The OpAL cross-lingual run 3 obtaining the highest F score for the non-agreed measures and ranking second according to the agreed measures. On the other hand, we can deduce that LSA-based method to determine topic-related words is not enough to perform this task. The terms obtained by employing this method are correct and

[9] http://aye.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html

useful, however, as future work our purpose is to use language models, to better account for the language variability. Finally, we understand that correference or temporal resolution systems do not improve the performance of OQA, and as a consequence we will study the performance of other correference and temporal resolution systems in order to check if the technique is not enough mature or if other systems can bring added value to this task.

8. 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/ 2010/288).

We would also like to thank our colleagues - Dr. José Manuel Gómez and Javier Fernández - for the help in the snippet retrieval step, using JIRS.

9. REFERENCES

[1] Seki, Y., Ku, L.-W., Sun, L., Chen, H.-H., Kando, N. Overview of Multilingual Opinion Analysis Task at NTCIR-8 – A Step Toward Cross-lingual Opinion Analysis. In Proceedings of NTCIR 8, 2010.

[2] Cardie. C. Wiebe. J. Wilson. T. 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.

[3] Yu. H. Hatzivassiloglou. V. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions. In Proceedings of EMNLP-03.

[4] Kim, S. M. and E.H. Hovy. 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.

[5] 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.

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

[7] Wilson, T., J. Wiebe, and Hoffmann, P. 2005. Recognizing 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.

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

[9] Wenjie, L., Ouyang, Y., Hu, Y., Wei, F. PolyU at TAC 2008. In Proceedings of Text Analysis Conference, at the joint annual meeting of TAC and TREC. Gaithersburg, Maryland, USA.

[10] Li, F., Zheng, Z.,Yang T., Bu, F., Ge, R., Zhu, X., Zhang, X., and Huang, M. THU QUANTA at TAC 2008. QA and RTE track. In Proceedings of Text Analysis Conference 2008, at the Joint Annual Meeting of TAC and TREC. Gaithersburg, Maryland, USA.

[11] Gómez, J.M., Montes-y-Gómez, M., Sanchis, E., Rosso, P. (2005). A Passage Retrieval System for Multilingual Question Answering. In Text, Speech and Dialogue: 8th International Conference, TSD 2005 (pp 443-450). Lecture Notes in Computer Science, 3658/2005. Springer. Heidelberg, Berlin.

[12] Saquete, E., Muñoz, R., Martínez-Barco, P. Event Ordering using TERSEO system. Data Knowledge and Engineering, 2006.

[13] Balahur, A. and Montoyo, A. 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, 2008.

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

[15] Wilson, T., Wiebe, J., and Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of HLT-EMNLP 2005.