Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 267-271 recibido 21-06-10 revisado 05-07-10 aceptado 06-07-10

IBEREVAL OM: Mining Opinions from the new textual genres

IBEREVAL OM: Minería de opiniones en los nuevos géneros textuales

Alexandra Balahur
University of Alicante, Department of
Language and Computing Systems
abalahur@dlsi.ua.es

Ester Boldrini
University of Alicante, Department of
Language and Computing Systems
eboldrini@dlsi.ua.es

Andrés Montoyo
University of Alicante, Department of
Language and Computing Systems
montoyo@dlsi.ua.es

Patricio Martínez-Barco
University of Alicante, Department of
Language and Computing Systems
patricio@dlsi.ua.es

Abstract: The increasing amount of subjective data on the Web is creating the need to develop effective Question Answering systems able to discriminate such information from factual data, and subsequently process it with specific methods. The participants in the IBEREVAL OM tasks will be given a set of opinion questions (in Spanish and English). Optionally, they will also be able to receive the same set of opinion questions, in which the source, target and expected polarity, as well as the time span the question is referring to are given. They will also be provided with a collection of blog posts, extracted using the Technorati blog search engine (in Spanish and English), in which the answers to the opinion questions should be found

The gold standard for this blog posts collection will previously be annotated using the EmotiBlog scheme, by a number of 3 annotators. The EmotiBlog corpus and the set of questions presented in (Balahur et al., 2009) – in their present state will be provided for system training. The participants will be able to participate in two subtasks : 1) in the first one, they will be asked to provide the list of answers to each of the questions (in the same language as the questions, or in the other language); 2) in the second one, they will be asked to provide a summary of the question answers – the top x% of the most important answers, in a non-redundant manner. The Gold Standard for the summaries will be automatically extracted from the manual annotations, taking into account the “intensity” parameter of the opinions expressed.

Resumen: Con el grande aumento de la información sujetiva en la Web, hay una importante necesidad de desarrollar sistemas de Question Answering que sen eficientes y capaces de discriminar entre datos objetivos y sujetivos. Los participantes tendrán una colección de preguntas de opinión (Español e Inglés) en las cuales se deberán encontrar las respuestas. El Gold Standard será anotado previamente con el esquema de anotación EmotiBlog por 3 anotadores. El corpus EmotiBlog y la colección de preguntas presentados en (Balahur et al. 2009) se pondrá a disposición para el entrenamiento del sistema. Los participantes deberán devolver un listado de respuestas para cada una de las preguntas, (en el mismo idioma que la pregunta o en otro), un resumen de las respuestas –de las x% de las respuestas más importantes, de una manera no redundante , el Gold Standard para los resúmenes será extraído automáticamente de las anotaciones manuales teniendo en consideración el parámetro de “intensidad” de la opinión expresada.

1 Introduction

Relevant surveys, such as the one carried out by the Technorati blog search engine, entitled “State of the Blogosphere 2009”[1] demonstrated the urgent need to develop Natural Language (NLP) tools able to deal with subjective data, which is in constant increase. Such data is

[1] http://technorati.com/

highly relevant, as it consists of genuine and unbiased information provided directly by the people involved in it. Given these properties, this data can be used for many studies with practical social and economic applications, focused on the benefit of the entire community. The Natural Language Processing (NLP) task dealing with the treatment of subjective data is called Opinion Mining (OM). It is focused on giving the users the appropriate instruments to efficiently access this subjective data, through their queries. Much research has been carried out focused on retrieving subjective information (Cardie et al., 2003, Yu and Hatzivassiloglou, 2003, Kim and Hovy, 2005). (Kim and Hovy, 2005) identified opinion holders, which are an important factor to be taken into consideration in the context of opinión questions (Balahur et al, 2009). In recent years, due to the increasing importance of subjective data present on the Internet, we have witnessed a growth in interest for performing NLP research focused on the development of opinion-related systems. While some of the efforts concentrate on the development of opinion-specific techniques – e.g. for opinion extraction and classification – others concentrate on integrating opinion mining into more complex systems – e.g. opinion QA systems. Special benchmark competitions have been organized with the aim of supporting this new line of research. 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. Example of the participating systems can be the Alyssa (Shen et al, 2007) which uses a Support Vector Machines (SVM) classifier trained on the MPQA corpus (Wiebe, Wilson and Cardie, 2005), English NTCIR[3] data and rules based on the subjectivity lexicon (Wilson, Wiebe and Hoffman, 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 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. For the English monolingual subtask in NTCIR 8 MOAT, 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 the question refers to. 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). 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).

2 Motivation

Having analysed the tasks, which are already set up, we can deduce that one of the first problems is that all of them or the majority are designed for English, thus there is a significant

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

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

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

lack of tools and evaluation benchmarks for other languages. The unique attempt to built up a task designed for language other than English can be seen in the NTCIR 8 MOAT in which we can see an approach to Chinese.

An additional challenge of OM is marked by the lack of annotated corpora (form the new textual genres) in languages other than English. One of the rare examples can be the EmotiBlog corpus (Boldrini et al, 2009) a collection of blog posts in English, Italian and Spanish labelled with the EmotiBlog annotation schema. Previous works (Wiebe, Wilson and Cardie, 2005, Pang and Lee, 2008) have carried out research but at a coarse-grained level. Contrary to this approach, the purpose of EmotiBlog is to respond to the need for multilingual resources, of different domains and labelled at a fine-grained level. Apart from this rich analysis of text, possible through its structure, EmotiBlog can also be employed to annotate text at a sentence and document level, depending on the needs. The added value of EmotiBlog is that previous systems cannot perform the joint tasks of topic/target/source opinion mining using more complex features. It is composed by:

- Objective speech: annotator's confidence (high, medium law), comment (if necessary), source (writer) and target (discourse topic);

In some cases writers use rhetoric strategies to state something that is apparently objective, but it in fact an indirect expression of a subjective point of view. In order to be able to contemplate these cases, we inserted in the model the following elements (for explanation, please see (Balahur and Montoyo, 2008, Balahur and Steinberger, 2009, Balahur et al., 2010):

- Reader Interpretation: annotator's confidence, comment, level, emotion, phenomenon, polarity, source and target. It is employed for capturing the impression/feeling/reaction the reader has when reading the text, by interpreting its meaning based on his personal beliefs and what s/he can affectively experience from reading the piece of text.
- Author Interpretation: annotator's confidence, comment, level, emotion, phenomenon, polarity, source and target. This element is used to understand what we can deduce from the author (politic orientation, preferences) thanks to the words and language s/he chooses. It is also a marker for bias introduced in a subtle way.

For both objective and subjective speech, the annotator has to specify the nature of the sentence s/he is labelling:

- Phenomenon: annotator's confidence, comment, type. This element explains the nature of the sentence we are labelling. They can be collocation, saying, slang, title, and rhetoric. A saying is a well-known and wise statement, which often has a meaning, different from the simple meaning of the words it contains[5]; while a collocation is a word or phrase, which is frequently used with another word or phrase, in a way that sounds correct to native speakers, but might not be expected from the individual words' meanings[6].

In case the annotator is labelling a subjective sentence, s/he will first label the entire sentence, underlining its nature, using the following tag:

- Subjective speech: annotator's confidence, comment, level, emotion, phenomenon, polarity, source and target.

In case of a subjective sentence, the annotator has to detect the elements, which give the subjectivity shadow to the discourse. *EmotiBlog* contemplates the ones below:

- Adjective/Adverbs: annotator's confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target.
- Verbs: annotator's confidence, comment, level, emotion, phenomenon, polarity, mode, source and target.
- Nouns: annotator's confidence, comment, level, emotion, phenomenon, modifier/not, polarity, and source.
- Anaphora: annotator's confidence, comment, type, source and target. This element underlines the correference phenomena at a cross-post level. Usually,

[5] Definition according to the Cambridge Advanced Learner's Dictionary

[6] Definition according to the Cambridge Advanced Learner's Dictionary

blog posts and their subsequent comments are similar to a multi-party conversation and thus this element can be useful to follow the discourse in case of multiple posts or when it is interrupted with other posts about a subtopic or related topic.

- Capital Letter: annotator's confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target. Bloggers generally produce a genuine and spontaneous language and it is frequent to find complete words that are meant as a sign of a special user attitude.
- Punctuation: annotator's confidence, comment, level, emotion, phenomenon, modifier/not, polarity, source and target. This phenomenon is similar to the previous one. An exceptional use of punctuation could mean a special feeling of the writer.
- Emotions: annotator's confidence, comment, accept, anger, anticipation, anxiety, etc.

Regarding the list of emotions employed, we grouped all sentiments into subgroups to facilitate the evaluation process. Emotions of the same subgroup will have less impact when calculating the inter-annotation agreement. In order to make this subdivision proper and effective, we were inspired by (Scherer, 2005). We started from this classification, grouping sentiments into positive and negative, and we also divided them as high/low power control, obstructive/conductive and active/passive. Further on, we distributed the sentiments within our list into the Scherer slots, creating other smaller categories included in the abovementioned general ones.

3 Proposal

IBEREVAL OM aims to be a framework for evaluating OM systems, under clearly defined settings, using as standards the definition of labels and annotations in EmotiBlog. Thus, it will allow the training of more complex features since the EmotiBlog annotation is fine-grained. Performing this task OM systems will be able to perform multilingual opinion retrieval for English-Spanish-Italian as can be seen in (Balahur et al. 2009 and Balahur et al, 2010 a and b)

The participants will be given:

- A set of opinion questions (in Spanish and English)
- Optionally, the same set of opinion questions, in which the source, target and expected polarity, as well as the time span the question is referring to are given
- A collection blog posts, extracted with using the Technorati blog search engine (in Spanish and English), in which the answers to the opinion questions should be found

The gold standard(for the corpus of blogs extracted with Technorati) will previously be annotated using the EmotiBlog scheme, by a number of 3 annotators

The EmotiBlog corpus and the set of questions presented in (Balahur et al., 2009)– in its present state, will be provided for system training

The participants will be able to participate in two subtasks:

1. To provide the list of answers to each of the questions (in the same language as the questions, or in the other language)
2. To provide a summary of the question answers – the top x% of the most important answers, in a non-redundant manner.

The Gold Standard for the summaries will be automatically extracted from the manual annotations, taking into account the "intensity" parameter of the opinions expressed.

4 Levels of contributions

IBEREVAL OM purpose is to establish the benchmark for tasks definition and understanding. It will also give the opportunity to build up comparable systems. IBEREVAL OM will be the scenario to offer the systems the possibility to jointly perform some of the tasks and to determine a holistic approach to effective solve the OM process. It will also establish a multilingual set for innovative tasks such as Opinionated Information Retrieval and Opinionated Question Answering.

Acknowledgements

This evaluation task proposal has been partially supported by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), and Conselleria d'Educació - Generalitat Valenciana (grant no. PROMETEO/2009/119 and ACOMP/2010/288.

References

Balahur, A. and Montoyo, A. 2008. *An incremental multilingual approach to forming a culture dependent emotion triggers lexical database*. In Proceedings of the Conference of Terminology and Knowledge Engineering (TKE 2008).

Balahur, A, Steinberg, R., *Rethinking Sentiment Analysis in the News:from Theory to Practice and back.* In Proceedings of WOMSA 2008, Seville, Spain.

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. 2009. *Opinion and Generic Question Answering systems: a performance analysis.* In Proceedings of ACL, 2009, Singapore.

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. 2009. *Opinion Question Answering: Towards a Unified Approach*. To appear in proceedings of the ECAI conference.a

Balahur, A., Boldrini, E., Montoyo, A., Martínez-Barco, P. 2010. *A Unified Proposal for Factoid and Opinionated Question Answering.* To appear in proceedings of the COLING conference.b

Balahur A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J. 2010. *Sentiment Analysis in the News*. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010) Valletta, Malta.

Boldrini, E., Balahur, A. Martínez-Barco, P. and Montoyo A. 2009. *EmotiBlog: An Annotation Scheme for EmotionDetection and Analysis in Non-Traditional Textual Genres.* In Proceedings of the 5th Conference on Data Mining. Las Vegas, Nevada, USA.

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.

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.

Pang, B. and Lee, L. (2008). *Opinion mining and sentiment analysis.*

Shen. D. Leidner. J. Merkel. A. 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.

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

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

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

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 Human Language Technologies Conference/Conference on Empirical methods in Natural Language Processing (HLT/EMNLP), Vancouver, BC, Canada.