Using WordNet for Opinion Mining

Pavel Smrž
Faculty of Information Technology
Brno University of Technology
Božetěchova 2, 612 66 Brno, Czech Republic
smrz@fit.vutbr.cz

Abstract

This paper deals with lexical resources applied for opinion mining – the identification and extraction of opinions from free texts. Opinion mining comprises the segmentation of documents, passages, sentences, or phrases to objective (factual) and subjective parts, and the evaluation of the subjective attitude toward a given fact. We briefly introduce an automatic system that was designed to crawl various information sources available on the Web – newspapers, Internet blogs and forums – to collect and identify different opinions on a given topic and to report diversity of opinions across languages and countries. A special attention is paid to linguistic resources used, especially to wordnet extensions that play a crucial role in the identification of subjective expressions.

Introduction

Subjectivity identification is one of the hot topics of the current text analysis research. Its popularity is driven by the expectation that information retrieval and extraction, traditionally focused on the subject matter (factual content) of text, could also provide appropriate methods for distinguishing between factual and non-factual information, extracting the subjective part of the data and report and/or summarize different viewpoints. Such a solution would become a basis for many natural-language processing applications – attitude and feeling trackers in the news and other on-line information sources, multi-document summarization, intelligent information extraction, multi-perspective question answering etc.

The ultimate goal of our work is ambitious – to design and develop a system that will analyse various information sources – news streams, Internet blogs and forums – in several countries (languages), to identify and collect opinionated texts, and report the diversity of the opinions on the same issue across countries and languages. The system will be available both in the form of a professional application and in the form of web services open to the public.

1 Design of the system

The figure on the following page shows the architecture of the system. The proposed schema of the interaction with the system is as follows: A user formulates a query in a particular language and enters it through a web interface in his/her language. Very simple processing is performed that identifies the main focus (topic) of the query and region(s) of the interest.

Cross-language information retrieval (CLIR) is launched next. It takes advantage of the freely available web translational engines and bilingual dictionaries to search the web for documents relevant to the given topic.

Topic crawler extracts relevant documents from the Internet. They take advantage of the standard web search engines such as Google, Yahoo, etc., as well as specialized information sources accessible by means of RSS, Atom and other relevant protocols.

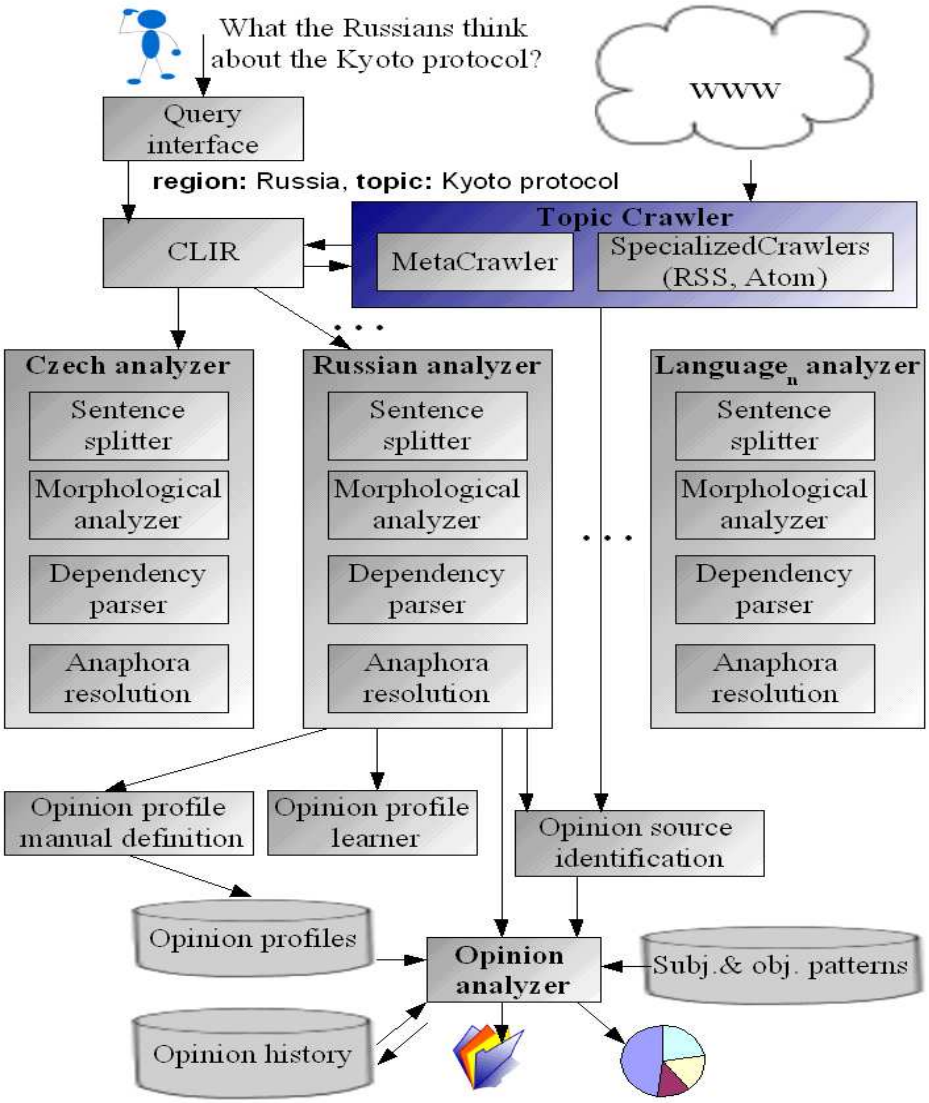
The language of the retrieved documents determines the next processing steps. Language analyzers are provided for the focused languages. This part of the system needs a special attention. Currently, the analyzers are available for a limited number of languages. For Czech all the analyzer components are at our disposal (the anaphora resolution module is rather rudimentary). Most of the modules are prepared for English and Russian.

The control module – Opinion analyzer – combines the results of the language analysis with the defined/extracted profiles and produces the final results. The component consults the identified sources of the opinion and filters information based on the user requirements (e.g. summarizes just the direct opinions and ignores all mediated ones). To extract the correct part of the relevant document, opinion analyzer puts into play the database of the subjective and objective patterns. The creation of the resource is discussed in the next section.

2 Using WordNet for subjectivity clue mining

It has been shown earlier (Riloff et al. (2003)) that the clues of subjectivity can be learned automatically from relevant corpora. However, it is not easy to get hold of such resources, especially in the multilingual environment we aim at. Document-level systems can take advantage of reviews, editorials, etc. available on the Internet. Phrase-level systems need either a large annotated corpus which is very difficult to obtain for several languages, or a comprehensive list of seed words that can initialize the learning process. We have chosen the second alternative for our work. The availability of the English wordnet that covers the entire core lexicon needed for our experiments as well as the wordnets for other

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 333–335.



languages that can be linked to the English one provided the basis for the multilingual subjectivity clue mining.

We have started with a short list of English words describing cognitive states, attitudes and feelings. The synsets containing the words were selected from Princeton WordNet 2.1. If a sense of the given word was not relevant for our task, the synset had been manually eliminated from the future processing. Next, we explored hypernyms of the identified synsets. A simple statistics helped us to determine where it is appropriate to generalize given synsets and include the hypernym and all the co-hyponyms to our list as well.

Other wordnet relations needed more attention. Inspired by Strapparava and Valitutti (2004) and Grefenstette et al. (2004), we examined each particular relation whether it is subjectivity sensitive, i.e., if two synsets are related to each other and one of them is in our current list of subjectivity clues, the second should be considered too. Analysing the existing list (what are the most frequent relations between the words that are initially in it), we identified antonymy, similarity, see-also and attribute as the most reliable ones. Antonymy changes the polarity (positive vs. negative) of the clue (in most cases), similarity usually modifies the intensity of the clue (strong vs. weak). The characteristics of the words obtained from the other relations were set manually.

The existing wordnets (if available) and bilingual dictionaries were applied in the next step. The list of English seed words acquired by the procedure given above was automatically converted to the other languages and the subjectivity characteristics (polarity and intensity) were transferred as well. Of course, there are cases where such an automatic

approach generates errors (in a particular language the translation of an English word cannot be used in a subjective expression). However, as the list is taken as a seed lexicon in the next processing only, no special care is given to this situation.

3 Conclusions and future directions

The multilingual subjectivity clue-mining system described in the previous section is a crucial component of the opinion mining system we are working on. The Princeton wordnet as well as national wordnets available for the particular languages and linked to the English one present valuable resources for the seed subjectivity lexicons needed in the extraction algorithm. Without them, the process of building a multilingual opinion extraction system would need much more annotated data for subjectivity learning methods.

Wordnet is regularly used for automatic word-sense disambiguation (WSD). The opinion mining system can take advantage of WSD, especially in the phase of the extraction of opinionated passages supporting a view of the author (or a referred information source). Our future research will therefore focus on a combination of the implemented clue extraction system with the wordnet-based WSD that should increase the accuracy of the tool.

Acknowledgements

This paper resulted from discussions with colleagues from the proposed project mentioned in the paper. Thanks belong to C. H. A. (Kees) Koster from the Institute of Computer Sciences, Radboud University Nijmegen, the Netherlands, Eneko Agirre from the IXA group, University of the Basque Country, Spain, Emanuele Pianta from the Istituto Trentino di Cultura – Centro per la Ricerca Scientifica e Tecnologica, Italy, Adam Przepiórkowski from the Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland, Irina Azarova from the Saint-Petersburg University and Ideograph, Russia, and Anna Sinopalnikova from the Faculty of Information Technologies, Brno University of Technology, Czech Republic.

References

Grefenstette G., Qu Y., Evans D. A., and Shanahan J. G. (2004) *Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes*. In: Proceedings of the AAAI SS on Exploring Attitude and Affect in Text: Theories and Application. Stanford University.

Riloff E., Wiebe J., and Wilson T. (2003) *Learning subjective nouns using extraction pattern bootstrapping.* In: Proceedings of the CoNLL-03 Conference".

Strapparava C. and Valitutti A. (2004) *WordNet-affect: An affective extension of WordNet*, Proceedings of LREC, Lisabon, Portugal.