OpinionFinder: A system for subjectivity analysis

Theresa Wilson[‡], Paul Hoffmann[‡], Swapna Somasundaran[†], Jason Kessler[†], Janyce Wiebe[†‡], Yejin Choi[§], Claire Cardie[§], Ellen Riloff[*], Siddharth Patwardhan[*]

[‡]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260
[†]Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260
[§]Department of Computer Science, Cornell University, Ithaca, NY 14853
[*]School of Computing, University of Utah, Salt Lake City, UT 84112

{twilson,hoffmanp,swapna,jsk44,wiebe}@cs.pitt.edu,
{ychoi,cardie}@cs.cornell.edu, {riloff,sidd}@cs.utah.edu

1 Introduction

OpinionFinder is a system that performs *subjectivity analysis*, automatically identifying when opinions, sentiments, speculations, and other *private states* are present in text. Specifically, OpinionFinder aims to identify *subjective* sentences and to mark various aspects of the subjectivity in these sentences, including the *source* (holder) of the subjectivity and words that are included in phrases expressing positive or negative sentiments.

Our goal with OpinionFinder is to develop a system capable of supporting other Natural Language Processing (NLP) applications by providing them with information about the subjectivity in documents. Of particular interest are question answering systems that focus on being able to answer opinion-oriented questions, such as the following:

> How is Bush's decision not to ratify the Kyoto Protocol looked upon by Japan and other US allies?
>
> How do the Chinese regard the human rights record of the United States?

To answer these types of questions, a system needs to be able to identify when opinions are expressed in text and who is expressing them. Other applications that would benefit from knowledge of subjective language include systems that summarize the various viewpoints in a document or that mine product reviews. Even typical fact-oriented applications, such as information extraction, can benefit from subjectivity analysis by filtering out opinionated sentences (Riloff et al., 2005).

2 OpinionFinder

OpinionFinder runs in two modes, batch and interactive. Document processing is largely the same for both modes. In batch mode, OpinionFinder takes a list of documents to process. Interactive mode provides a front-end that allows a user to query on-line news sources for documents to process.

2.1 System Architecture Overview

OpinionFinder operates as one large pipeline. Conceptually, the pipeline can be divided into two parts. The first part performs mostly general purpose document processing (e.g., tokenization and part-of-speech tagging). The second part performs the subjectivity analysis. The results of the subjectivity analysis are returned to the user in the form of SGML/XML markup of the original documents.

2.2 Document Processing

For general document processing, OpinionFinder first runs the Sundance partial parser (Riloff and Phillips, 2004) to provide semantic class tags, identify Named Entities, and match extraction patterns that correspond to subjective language (Riloff and Wiebe, 2003). Next, OpenNLP[1] 1.1.0 is used to tokenize, sentence split, and part-of-speech tag the data, and the Abney stemmer[2] is used to stem. In batch mode, OpinionFinder parses the data again, this time to obtain constituency parse trees (Collins, 1997), which are then converted to dependency parse trees (Xia and Palmer, 2001). Currently, this stage is only

[1] http://opennlp.sourceforge.net/
[2] SCOL version 1g available at http://www.vinartus.net/spa/

Proceedings of HLT/EMNLP 2005 Demonstration Abstracts, pages 34–35,
Vancouver, October 2005.

available for batch mode processing due to the time required for parsing. Finally, a clue-finder is run to identify words and phrases from a large subjective language lexicon.

2.3 Subjectivity Analysis

The subjectivity analysis has four components.

2.3.1 Subjective Sentence Classification

The first component is a Naive Bayes classifier that distinguishes between subjective and objective sentences using a variety of lexical and contextual features (Wiebe and Riloff, 2005; Riloff and Wiebe, 2003). The classifier is trained using subjective and objective sentences, which are automatically generated from a large corpus of unannotated data by two high-precision, rule-based classifiers.

2.3.2 Speech Events and Direct Subjective Expression Classification

The second component identifies speech events (e.g., "said," "according to") and direct subjective expressions (e.g., "fears," "is happy"). Speech events include both speaking and writing events. Direct subjective expressions are words or phrases where an opinion, emotion, sentiment, etc. is directly described. A high-precision, rule-based classifier is used to identify these expressions.

2.3.3 Opinion Source Identification

The third component is a source identifier that combines a Conditional Random Field sequence tagging model (Lafferty et al., 2001) and extraction pattern learning (Riloff, 1996) to identify the sources of speech events and subjective expressions (Choi et al., 2005). The source of a speech event is the speaker; the source of a subjective expression is the experiencer of the private state. The source identifier is trained on the MPQA Opinion Corpus[3] using a variety of features. Because the source identifier relies on dependency parse information, it is currently only available in batch mode.

2.3.4 Sentiment Expression Classification

The final component uses two classifiers to identify words contained in phrases that express positive or negative sentiments (Wilson et al., 2005). The first classifier focuses on identifying sentiment expressions. The second classifier takes the sentiment expressions and identifies those that are positive and negative. Both classifiers were developed using BoosTexter (Schapire and Singer, 2000) and trained on the MPQA Corpus.

[3]The MPQA Opinion Corpus can be freely obtained at http://nrrc.mitre.org/NRRC/publications.htm.

3 Related Work

Please see (Wiebe and Riloff, 2005; Choi et al., 2005; Wilson et al., 2005) for discussions of related work in automatic opinion and sentiment analysis.

4 Acknowledgments

This work was supported by the Advanced Research and Development Activity (ARDA), by the NSF under grants IIS-0208028, IIS-0208798 and IIS-0208985, and by the Xerox Foundation.

References

Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *HLT/EMNLP 2005*.

M. Collins. 1997. Three generative, lexicalised models for statistical parsing. In *ACL-1997*.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML-2001*.

E. Riloff and W. Phillips. 2004. An Introduction to the Sundance and AutoSlog Systems. Technical Report UUCS-04-015, School of Computing, University of Utah.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *EMNLP-2003*.

E. Riloff, J. Wiebe, and W. Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *AAAI-2005*.

E. Riloff. 1996. An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *Artificial Intelligence*, 85:101–134.

R. E. Schapire and Y. Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

J. Wiebe and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing-2005*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP 2005*.

F. Xia and M. Palmer. 2001. Converting dependency structures to phrase structures. In *HLT-2001*.