Grammar-based approaches to opinion mining

Asad B. Sayeed, Ph.D.
Multimodel Computing and Innovation Cluster of Excellence
Saarland University
66143 Saarbrücken, Saarland, Germany
asayeed@coli.uni-sb.de
http://www.coli.uni-saarland.de/~asayeed

Course type Advanced

Abstract This course will present recent work in the growing area of grammar-based approaches to opinion-mining, which are designed to handle genres for which opinions can be found at a granularity finer than sentence-level. These genres are associated with applications such as market prediction and corpus-based social science. We will cover existing training corpora, such as the MPQA and the recent JDPA, which have fine-grained annotations; we will also cover specialized corpus development, particularly through crowdsourcing. In terms of techniques, we will cover sequence models (labelling input strings by opinion weight), graph-based models (labelling grammatical structures), and we will refresh or introduce recent relevant advances in machine learning. Knowledge of basic machine learning and dependency parsing is required, but knowledge of opinion mining is not. Relevant references are taken from very recent experimental and survey papers.

1 Introduction

We are proposing a course in grammatical approaches to opinion mining. By this, we mean approaches that make use of the fine details of formal syntactic and semantic structure to identify or classify text that represents opinions or sentiments expressed about given topics. In this document, we describe some of the areas we will discuss in the course and provide some representative citations.

Opinion mining, also known as sentiment analysis (survey papers Pang and Lee, 2008; Liu, 2010), is a relatively recent area of research in natural language processing. It has grown very quickly as a research area, partly for economic reasons, and it has been developing around a small number of basic approaches. These approaches include "bag-of-words" models or sequence models (e.g. HMMs), but these are not necessarily appropriate for all circumstances, particularly in genres where many potential opinion-statements can be identified in the same stretch of text.

This problem is particularly relevant in the expansion of sentiment analysis techniques to areas such as market prediction (Bollen et al., 2010) and corpus-based social science (Sayeed et al., 2010b; Tsui et al., 2009). In these areas, it is not enough to predict or detect opinions in predefined areas of text or even to mine for the locations of opinions in large corpora, but it is necessary to be able to connect opinions across documents and to reconstruct the social networks that underlie social trends. Furthermore, it must be possible to do this in text that can have an arbitrary number of opinions intertwined in ways that go beyond the base case of product review text. This requires both additional consideration of the perspective of the user and attention to the finer-grained details of sentiment expression.

This advanced course will focus on (1) the challenges surrounding opinion mining at this level, (2) existing corpus resources and corpus development issues, (3) sequence-based techniques for sentiment-related grammatical structure identification, and (4) structure-based techniques.

2 Challenges

In applications of opinion mining that attempt to track the relationshp of opinions in a corpus to external trends such as stock market prediction (Bollen et al., 2010), the actual opinionated state of the source of the opinion not only matters less, but is hard to gauge. Furthermore, there are particular actions or statements (Austin, 1962) that have an effect on the real world, and could be said by implication to express an opinion. Buying and selling in certain contexts implies confidence in a product or service. We can refer to this as "pragmatic opinion" (Somasundaran and Wiebe, 2009).

For example, consider the following sentence from a major information technology (IT) business journal:

> Lloyd Hession, chief security officer at BT Radianz in New York, said that virtualization also opens up a slew of potential network access control issues.

Consider that there are a panoply of potential opinion holders (*sources*) in this sentence. There are also a number of mentioned opinion topics (*targets*). The actual opinions themselves are heavily conditioned on the perspective of the reader: what words are relevant to the reader's perspective and interests, and in what relation do they stand with the other words in the sentence?

The challenge in identifying reader-relevant opinion text at this level requires the use of indirect means. We can treat this as a search for *(source, target, opinion)* triples. A major piece of evidence are the (often semantically non-compositional) relationships between words in the sentence, which can be viewed as observations of latent pragmatic variables. This course will focus on how to use these relationships to guess the true values of the variables.

3 Corpus resources and development

Most work in this area depends on annotated corpora for development and model training. There have been recent efforts to produce publicly available corpus resources. The most established one is the Multi-Perspective Question-Answering (MPQA) newswire corpus (Wilson and Wiebe, 2005). A more recently available collection is the J. D. Power & Associates (JDPA) automotive review blog post (Kessler et al., 2010) corpus. Both contain sub-sentence annotations of sentiment-bearing language as text spans. In some cases, they also include links to within-sentence targets:

> That was the moment at which the fabric of compassion tore, and worlds cracked apart; when **the contrast and conflict of civilisational values** became so great as to *remove any sense of common ground* - even on which to do battle.

The italicised portion is annotated as conveying a negative sentiment towards the bolded target (MPQA).

This course will present some of the prerequisites for understanding and using the MPQA and the JDPA and for creating similar resources. The increasingly popular area of crowdsourcing also permits the development of finely-detailed sentiment resources, so we will discuss recent work on how to leverage crowdsourcing and user interface design for custom resource development (Sayeed et al., 2010a).

4 Sequence-based techniques

By far the largest selection of technologies for exploiting grammar in sentiment analysis come from the use of HMM- or CRF-type sequence modeling, and consequently this will be a major component of the course. This type of machine learning uses syntactic and other features as binary-valued functions in learning to label windows of text.

Examples include work such as Bethard et al. (2004) who use semantic role labelling, syntactic information, and lexical information as input to an SVM classifier that detects sentential complements that contain opinion propositions ("I think that **the fish tastes bad**"). Choi et al. (2005) experiment with extraction pattern recognition and conditional random field (CRF) based methods to recover opinion source information when the opinions themselves have already been marked in the MPQA corpus. Choi et al. (2006) use a CRF-based method alongside semantic role labelling to extract sources and opinions at the same time. Kim and Hovy (2006) make use of semantic frames and semantic role labelling to identify sources and targets; their frames come from an existing frame database (FrameNet).

More recently, Jakob and Gurevych (2010) perform target identification from known opinion spans within a sentence using paths extracted from dependency parses. They use the single shortest path from a target expression to an opinion span as a binary feature in a CRF model that labels words with their status as a word that belongs to a target expression.

5 Structure-based techniques

By "structure-based", we mean graph-based approaches that do not rely on direct labelling of strings in the input text. This generally means learning over dependency parse structures: classifying components of the sequence. This is the leading edge of this field, to which we will devote the final part of the course.

Examples of this work include Qiu et al. (2011), who use predefined heuristics over dependency parses to identify both targets and opinion keywords. Other work includes Nakagawa et al. (2010), who use the recent Bayesian inference technique of factor graph modeling over a dependency parse formalism for opinion polarity (positive/negative) classification. Sayeed et al. (2012) also use factor graph modeling (McCallum et al., 2009) for opinion text identification.

References

Austin, J. L. (1962). *How to do things with words*. Clarendon, Cambridge, Mass.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*.

Bollen, J., Mao, H., and Zeng, X.-J. (2010). Twitter mood predicts the stock market. *CoRR*, abs/1010.3003.

Choi, Y., Breck, E., and Cardie, C. (2006). Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jakob, N. and Gurevych, I. (2010). Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *EMNLP*.

Kessler, J. S., Eckert, M., Clark, L., and Nicolov, N. (2010). The 2010 ICWSM JDPA sentment corpus for the automotive domain. In *4th Int'l AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

Kim, S.-M. and Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *SST '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Liu, B. (2010). Sentiment analysis and subjectivity. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

McCallum, A., Schultz, K., and Singh, S. (2009). Factorie: Probabilistic programming via imperatively defined factor graphs. In *Neural Information Processing Systems (NIPS)*.

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *HLT-NAACL*.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2).

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9—27.

Sayeed, A., Boyd-Graber, J., Rusk, B., and Weinberg, A. (2012). Grammatical structures for word-level sentiment detection. In *Proceedings of NAACL 2012*, Montréal, Canada.

Sayeed, A. B., Meyer, T. J., Nguyen, H. C., Buzek, O., and Weinberg, A. (2010a). Crowdsourcing the evaluation of a domain-adapted named entity recognition system. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Sayeed, A. B., Nguyen, H. C., Meyer, T. J., and Weinberg, A. (2010b). Expresses-an-opinion-about: using corpus statistics in an information extraction approach to opinion mining. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10.

Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, ACL '09.

Tsui, C.-J., Wang, P., Fleischmann, K., Oard, D., and Sayeed, A. (2009). Understanding IT innovations by computational analysis of discourse. In *International conference on information systems*.

Wilson, T. and Wiebe, J. (2005). Annotating attributions and private states. In *CorpusAnno '05: Proceedings of the Workshop on Frontiers in Corpus Annotations II*, Morristown, NJ, USA. Association for Computational Linguistics.

Tentative outline

- Day 1: Context and foundations.
  - Sentiment analysis background: movie/product review classification, opinion stances in debates (30 minutes)
  - Challenges of perspective, pragmatic factors in opinion (30 minutes).
  - Introduction to corpus-based social science (30 minutes).
- Day 2: Corpus resources and construction.
  - Corpus requirements for fine-grained sentiment analysis (20 minutes)
  - JDPA and MPQA: uses, design choices, differences (35 minutes)
  - Crowdsourcing for sentiment analysis (35 minutes)
- Day 3: Machine learning.
  - Refresher: HMMs, CRFs, Bayesian inference (45 minutes)
  - Graphical models, factor graphs, tools (45 minutes)
- Day 4: Sequence-based techniques.
  - Opinion source identification techniques (30 minutes)
  - Opinion target challenges (30 minutes)
  - Polarity detection and classificatino (30 minutes)
- Day 5: Structure-based techniques.
  - Parser technology for sentiment analysis (20 minutes)
  - Rule/heuristic-based techniques (25 minutes)
  - Machine learning over structures (45 minutes)

Prerequisites

This is an advanced course, so attendees are assumed to have already acquired the basic foundations of statistical NLP. Knowledge/experience with dependency parsing is required, including familiarity with typical treebank tags for English (ie, Penn Treebank) and English-language syntactic dependencies. Knowledge of basic machine learning techniques and concepts is also required (e.g. probability theory, support vector machines, HMMs).

Prior experience with sentiment analysis/opinion mining is **not** expected.

Funding

Organizer is presently located within Germany (Saarbrücken). Some travel support may be available from the Multimodal Computing and Innovation Cluster of Excellence.