CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging

Alina Andreevskaia
Concordia University
1455 de Maisonneuve Blvd.
Montreal, Canada
andreev@cs.concordia.ca

Sabine Bergler
Concordia University
1455 de Maisonneuve Blvd.
Montreal, Canada
bergler@cs.concordia.ca

Abstract

For the Affective Text task at Semeval-1/Senseval-4, the CLaC team compared a knowledge-based, domain-independent approach and a standard, statistical machine learning approach to ternary sentiment annotation of news headlines. In this paper we describe the two systems submitted to the competition and evaluate their results. We show that the knowledge-based unsupervised method achieves high accuracy and precision but low recall, while supervised statistical approach trained on small amount of in-domain data provides relatively high recall at the cost of low precision.

1 Introduction

Sentiment tagging of short text spans — sentences, headlines, or clauses — poses considerable challenges for automatic systems due to the scarcity of sentiment clues in these units: sometimes, the decision about the text span sentiment has to be based on just a single sentiment clue and the cost of every error is high. This is particularly true for headlines, which are typically very short. Therefore, an ideal system for sentiment tagging of headlines has to use a large set of features with dependable sentiment annotations and to be able to reliably deduce the sentiment of the headline from the sentiment of its components.

The valence labeling subtask of the Affective Text task requires ternary — positive vs. negative vs. neutral — classification of headlines. While such categorization at the sentence level remains relatively unexplored[1], the two related sentence-level, binary classification tasks — positive vs. negative and subjective vs. objective — have attracted considerable attention in the recent years (Hu and Liu, 2004; Kim and Hovy, 2005; Riloff et al., 2006; Turney and Littman, 2003; Yu and Hatzivassiloglou, 2003). Unsupervised knowledge-based methods are the preferred approach to classification of sentences into positive and negative, mostly due to the lack of adequate amounts of labeled training data (Gamon and Aue, 2005). These approaches rely on presence and scores of sentiment-bearing words that have been acquired from dictionaries (Kim and Hovy, 2005) or corpora (Yu and Hatzivassiloglou, 2003). Their accuracy on news sentences is between 65 and 68%.

Sentence-level subjectivity detection, where training data is easier to obtain than for positive vs. negative classification, has been successfully performed using supervised statistical methods alone (Pang and Lee, 2004) or in combination with a knowledge-based approach (Riloff et al., 2006).

Since the extant literature does not provide clear evidence for the choice between supervised machine learning methods and unsupervised knowledge-based approaches for the task of ternary sentiment classification of sentences or headlines, we developed two systems for the Affective Text task at SemEval-2007. The first system (CLaC) relies on the knowledge-rich approach that takes into consid-

[1] To our knowledge, the only work that attempted such classification at the sentence level is (Gamon and Aue, 2005) that classified product reviews.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 117–120,
Prague, June 2007. ©2007 Association for Computational Linguistics

eration multiple clues, such as a list of sentiment-bearing unigrams and valence shifters, and makes use of sentence structure in order to combine these clues into an overall sentiment of the headline. The second system (CLaC-NB) explores the potential of a statistical method trained on a small amount of manually labeled news headlines and sentences.

2 CLaC System: Syntax-Aware Dictionary-Based Approach

The CLaC system relies on a knowledge-based, domain-independent, unsupervised approach to headline sentiment detection and scoring. The system uses three main knowledge inputs: a list of sentiment-bearing unigrams, a list of valence shifters (Polanyi and Zaenen, 2006), and a set of rules that define the scope and results of combination of sentiment-bearing words with valence shifters.

2.1 List of sentiment-bearing words

The unigrams used for sentence/headline classification were learned from WordNet (Fellbaum, 1998) dictionary entries using the STEP system described in (Andreevskaia and Bergler, 2006b). In order to take advantage of the special properties of WordNet glosses and relations, we developed a system that used the human-annotated adjectives from (Hatzivassiloglou and McKeown, 1997) as a seed list and learned additional unigrams from WordNet synsets and glosses. The STEP algorithm starts with a small set of manually annotated seed words that is expanded using synonymy and antonymy relations in WordNet. Then the system searches all WordNet glosses and selects the synsets that contain sentiment-bearing words from the expanded seed list in their glosses. In order to eliminate errors produced by part-of-speech ambiguity of some of the seed words, the glosses are processed by Brill's part-of-speech tagger (Brill, 1995) and only the seed words with matching part-of-speech tags are considered. Headwords with sentiment-bearing seed words in their definitions are then added to the positive or negative categories depending on the seed-word sentiment. Finally, words that were assigned contradicting — positive and negative — sentiment within the same run were eliminated. The average accuracy of 60 runs with non-intersecting seed lists when compared to General Inquirer (Stone et al., 1966) was 74%. In order to improve the list coverage, the words annotated as "Positiv" or "Negativ" in the General Inquirer that were not picked up by STEP were added to the final list.

Since sentiment-bearing words in English have different degree of centrality to the category of sentiment, we have constructed a measure of word centrality to the category of positive or negative sentiment described in our earlier work (Andreevskaia and Bergler, 2006a). The measure, termed Net Overlap Score (NOS), is based on the number of ties that connect a given word to other words in the category. The number of such ties is reflected in the number of times each word was retrieved from WordNet by multiple independent STEP runs with non-intersecting seed lists. This approach allowed us to assign NOSs to each unigram captured by multiple STEP runs. Only words with fuzzy membership score not equal to zero were retained in the list. The resulting list contained 10,809 sentiment-bearing words of different parts of speech.

2.2 Valence Shifters

The brevity of the headlines compared to typical news sentences[2] requires that the system is able to make a correct decision based on very few sentiment clues. Due to the scarcity of sentiment clues, the additional factors, such as presence of valence shifters, have a greater impact on the system performance on headlines than on sentences or texts, where impact of a single error can often be compensated by a number of other, correctly identified sentiment clues. For this reason, we complemented the system based on fuzzy score counts with the capability to discern and take into account some relevant elements of syntactic structure of sentences. We added to the system two components in order to enable this capability: (1) valence shifter handling rules and (2) parse tree analysis.

Valence shifters can be defined as words that modify the sentiment expressed by a sentiment-bearing word (Polanyi and Zaenen, 2006). The list of valence shifters used in our experiments was a com-

[2] An average length of a sentence in a news corpus is over 20 words, while the average length of headlines in the test corpus was only 7 words.

bination of (1) a list of common English negations, (2) a subset of the list of automatically obtained words with increase/decrease semantics, and (3) words picked up in manual annotation conducted for other research projects by two trained linguists. The full list consists of 490 words and expressions. Each entry in the list of valence shifters has an action and scope associated with it. The action and scope tags are used by special handling rules that enable our system to identify such words and phrases in the text and take them into account in sentence sentiment determination. In order to correctly determine the scope of valence shifters in a sentence, we introduced into the system the analysis of the parse trees produced by MiniPar (Lin, 1998).

As a result of this processing, every headline received a score according to the combined fuzzy NOS of its constituents. We then mapped this score, which ranged between -1.2 and 0.99, into the [-100, 100] scale as required by the competition organizers.

3 CLaC-NB System: Naïve Bayes

Supervised statistical methods have been very successful in sentiment tagging of texts and in subjectivity detection at sentence level: on movie review texts they reach an accuracy of 85-90% (Aue and Gamon, 2005; Pang and Lee, 2004) and up to 92% accuracy on classifying movie review snippets into subjective and objective using both Nave Bayes and SVM (Pang and Lee, 2004). These methods perform particularly well when a large volume of labeled data from the same domain as the test set is available for training (Aue and Gamon, 2005). The lack of sufficient data for training appears to be the main reason for the virtual absence of experiments with statistical classifiers in sentiment tagging at the sentence level.

In order to explore the potential of statistical approaches on sentiment classification of headlines, we implemented a basic Naïve Bayes classifier with smoothing using Lidstone's law of succession (with λ=0.1). No feature selection was performed.

The development set for the Affective Text task consisted of only 250 headlines, which is not sufficient for training of a statistical classifier. In order to increase the size of the training corpus, we augmented it with a balanced set of 900 manually annotated news sentences on a variety of topics extracted from the Canadian NewsStand database[3] and 200 headlines from different domains collected from Google News in January 2007[4].

The probabilities assigned by the classifier were mapped to [-100, 100] as follows: all negative headlines received a score of -100, all positive headlines +100, and neutral headlines 0.

4 Results and Discussion

Table 1 shows the results of the two CLaC systems for valence labeling subtask of Affective Text task compared to all participating systems average. The best subtask scores are highlighted in bold.

System	Pearson correl.	Acc.	Prec.	Rec.	F1
CLaC	**47.7**	**55.1**	**61.4**	9.2	16
CLaC-NB	25.4	31.2	31.2	**66.4**	**42**
Task average	33.2	44.7	44.85	29.6	23.7

Table 1: System results

The comparison between the two CLaC systems clearly demonstrates the relative advantages of the two approaches. The knowledge-based unsupervised system performed well above average on three main measures: the Pearson correlation between fine-grained sentiment assigned by CLaC system and the human annotation; the accuracy for ternary classification; and the precision of binary (positive vs. negative) classification. These results demonstrate that an accurately annotated list of sentiment-bearing words combined with sophisticated valence shifter handling produces acceptably accurate sentiment labels even for such difficult data as news headlines. This system, however, was not able to provide good recall.

On the contrary, supervised machine learning has very good recall, but low accuracy relative to the results of the unsupervised knowledge-based approach. This shortcoming could be in part reduced if more uniformly labeled headlines were available

[3] http://www.il.proquest.com/products_pq/descriptions/Canadian_newsstand.shtml

[4] The interannotator agreement for this data, as measured by Kappa, was 0.74.

for training. However, we can hardly expect large amounts of such manually annotated data to be handy in real-life situations.

5 Conclusions

The two CLaC systems that we submitted to the Affective Text task have tested the applicability of two main sentiment tagging approaches to news headlines annotation. The results of the two systems indicate that the knowledge-based unsupervised approach that relies on an automatically acquired list of sentiment-bearing unigrams and takes into account the combinatorial properties of valence shifters, can produce high quality sentiment annotations, but may miss many sentiment-laden headlines. On the other hand, supervised machine learning has good recall even with a relatively small training set, but its precision and accuracy are low. In our future work we will explore the potential of combining the two approaches in a single system in order to improve both recall and precision of sentiment annotation.

References

Alina Andreevskaia and Sabine Bergler. 2006a. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics*, Trento, IT.

Alina Andreevskaia and Sabine Bergler. 2006b. Semantic tag extraction from wordnet glosses. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, Genova, IT.

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *RANLP-05, the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Eric Brill. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*, 21(4).

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Michael Gamon and Anthony Aue. 2005. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, Ann Arbor, MI.

Vasileios Hatzivassiloglou and Kathleen B. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of ACL-97, 35nd Meeting of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain. ACL.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Tenth ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-04)*, pages 168–177.

Soo-Min Kim and Eduard Hovy. 2005. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of IJCNLP-05, the Second International Joint Conference on Natural Language Processing*, pages 61–66, Jeju Island, KR.

Dekang Lin. 1998. Dependency-based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*, pages 768–774, Granada, Spain.

Bo Pang and Lilian Lee. 2004. A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics*, pages 271–278.

Livia Polanyi and Annie Zaenen. 2006. Contextual Valence Shifters. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application*. Springer Verlag.

Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. 2006. Feature subsumption for opinion analysis. In *Proceedings of EMNLP-06, the Conference on Empirical Methods in Natural Language Processing*, pages 440–448, Sydney, AUS.

P. J. Stone, D.C. Dumphy, M.S. Smith, and D.M. Ogilvie. 1966. *The General Inquirer: a computer approach to content analysis*. M.I.T. studies in comparative politics. M.I.T. Press, Cambridge, MA.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21:315–346.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In Michael Collins and Mark Steedman, editors, *Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing*, pages 129–136, Sapporo, Japan.