SemEval-2010 Task 18: Disambiguating Sentiment Ambiguous Adjectives

Yunfang Wu
Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, China
wuyf@pku.edu.cn

Peng Jin
Laboratory of Intelligent Information Processing and Application, Leshan Normal University, China
jinp@lstc.edu.cn

Abstract

Sentiment ambiguous adjectives cause major difficulties for existing algorithms of sentiment analysis. We present an evaluation task designed to provide a framework for comparing different approaches in this problem. We define the task, describe the data creation, list the participating systems and discuss their results. There are 8 teams and 16 systems.

1 Introduction

In recent years, sentiment analysis has attracted considerable attention (Pang and Lee, 2008). It is the task of mining positive and negative opinions from natural language, which can be applied to many natural language processing tasks, such as document summarization and question answering. Previous work on this problem falls into three groups: opinion mining of documents, sentiment classification of sentences and polarity prediction of words. Sentiment analysis both at document and sentence level rely heavily on word level.

The most frequently explored task at word level is to determine the semantic orientation (SO) of words, in which most work centers on assigning a prior polarity to words or word senses in the lexicon out of context. However, for some words, the polarity varies strongly with context, making it hard to attach each to a specific sentiment category in the lexicon. For example, consider "low cost" versus "low salary". The word "low" has a positive orientation in the first case but a negative orientation in the second case.

Turney and Littman (2003) claimed that sentiment ambiguous words could not be avoided easily in a real-world application in the future research. But unfortunately, sentiment ambiguous words are discarded by most research concerning sentiment analysis (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Kim and Hovy, 2004). The exception work is Ding et al. (2008). They call these words as *context dependant opinions* and propose a holistic lexicon-based approach to solve this problem. The language they deal with is English.

The disambiguation of sentiment ambiguous words can also be considered as a problem of phrase-level sentiment analysis. Wilson et al. (2005) present a two-step process to recognize contextual polarity that employs machine learning and a variety of features. Takamura et al. (2006, 2007) propose latent variable model and lexical network to determine SO of phrases, focusing on "noun+adjective" pairs. Their experimental results suggest that the classification of pairs containing ambiguous adjectives is much harder than those with unambiguous adjectives.

The task 18 at SemEval 2010 provides a benchmark data set to encourage studies on this problem. This paper is organized as follows. Section 2 defines the task. Section 3 describes the data annotation. Section 4 gives a brief summary of 16 participating systems. Finally Section 5 draws conclusions.

2 Task Set up

2.1 Task description

In this task, we focus on 14 frequently used sentiment ambiguous adjectives in Chinese, which all have the meaning of measurement, as shown below.

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 81–85,
Uppsala, Sweden, 15-16 July 2010.

(1) Sentiment ambiguous adjectives(SAAs) ={大 da “large”, 多 duo “many”, 高 gao “high”, 厚 hou “thick”, 深 shen “deep”, 重 zhong “heavy”, 巨大 ju-da “huge”, 重大 zhong-da “great”, 小 xiao “small”, 少 shao “few”, 低 di “low”, 薄 bao “thin”, 浅 qian “shallow”, 轻 qing “light”}

These adjectives are neutral out of context, but when they co-occur with some target nouns, positive or negative emotion will be evoked. Although the number of such ambiguous adjectives is not large, they are frequently used in real text, especially in the texts expressing opinions and emotions.

The task is designed to automatically determine the SO of these sentiment ambiguous adjectives within context: positive or negative. For example, 高 gao “high” should be assigned as positive in 工资高 gong-zi-gao “salary is high” but negative in 价格高 jia-ge-gao “price is high”.

This task was carried out in an unsupervised setting. No training data was provided, but external resources are encouraged to use.

2.2 Data Creation

We collected data from two sources. The main part was extracted from Xinhua News Agency of Chinese Gigaword (Second Edition) released by LDC. The texts were automatically word-segmented and POS-tagged using the open software *ICTCLAS*[1]. In order to concentrate on the disambiguation of sentiment ambiguous adjectives, and reduce the noise introduced by the parser, we extracted sentences containing strings in pattern of (2), where the target nouns are modified by the adjectives in most cases.

(2) noun+adverb+adjective (adjective∈SAAs)
e.g. 成本/n 较/d 低/a cheng-ben-jiao-di
“the cost is low.”

Another small part of data was extracted from the Web. Using the search engine Google[2], we searched the queries as in (3):

(3) 很 hen “very”+ adjective (adjective∈SAAs)

From the returned snippets, we manually picked out some sentences that contain the strings of (2). Also, the sentences were automatically segmented and POS-tagged using *ICTCLAS*.

Sentiment ambiguous adjectives in the data were assigned as positive, negative or neutral, independently by two annotators. Since we focus on the distinction between positive and negative categories, the neutral instances were removed. The inter-annotator agreement is in a high level with a kappa of 0.91. After cases with disagreement were negotiated between the two annotators, a gold standard annotation was agreed upon. In total 2917 instances were provided as the test data in the task, and the number of sentences of per target adjective is listed in Table 2.

Evaluation was performed in micro accuracy and macro accuracy:

$$P_{mir} = \sum_{i=1}^{N} m_i / \sum_{i=1}^{N} n_i \quad (1)$$

$$P_{mar} = \sum_{i=1}^{N} P_i / N \qquad P_i = m_i / n_i \quad (2)$$

where N is the number of all target words, n_i is the number of all test instances for a specific word, and m_i is the number of correctly labeled instances.

2.3 Baseline

We group 14 sentiment ambiguous adjectives into two categories: positive-like adjectives and negative-like adjectives. The former has the connotation towards large measurement, whereas the latter towards small measurement.

(4) Positive-like adjectives (Pa) ={大 da “large”, 多 duo “many”, 高 gao “high”, 厚 hou “thick”, 深 shen “deep”, 重 zhong “heavy”, 巨大 ju-da “huge”, 重大 zhong-da “great”}

(5) Negative-like adjectives (Na) ={小 xiao “small”, 少 shao “few”, 低 di “low”, 薄 bao “thin”, 浅 qian “shallow”, 轻 qing “light”}

We conduct a baseline in the dataset. Not considering the context, assign all positive-like adjectives as positive and all negative-like adjectives as negative. The micro accuracy of the baseline is 61.20%.

The inter-annotator agreement of 0.91 can be considered as the upper bound of the dataset.

3 Systems and Results

We published firstly trial data and then test data. In total 11 different teams downloaded both the trial and test data. Finally 8 teams submitted their experimental results, including 16 systems.

[1] http://www.ictclas.org/.
[2] http://www.google.com/.

3.1 Results

Table 1 lists all systems' scores, ranked from best to worst performance measured by micro accuracy. To our surprise, the performance of different systems differs greatly. The micro accuracy of the best system is 94.20% that is 43.12% higher than the worst system. The accuracy of the best three systems is even higher than inter-annotator agreement. The performance of the worst system is only a little higher than random baseline, which is 50% when we randomly assign the SO of sentiment ambiguous adjectives.

Table 1: The scores of 16 systems

System	Micro Acc.(%)	Macro Acc.(%)
YSC-DSAA	94.20	92.93
HITSZ_CITYU_1	93.62	95.32
HITSZ_CITYU_2	93.32	95.79
Dsaa	88.07	86.20
OpAL	76.04	70.38
CityUHK4	72.47	69.80
CityUHK3	71.55	75.54
HITSZ_CITYU_3	66.58	62.94
QLK_DSAA_R	64.18	69.54
CityUHK2	62.63	60.85
CityUHK1	61.98	67.89
QLK_DSAA_NR	59.72	65.68
Twitter Sentiment	59.00	62.27
Twitter Sentiment_ext	56.77	61.09
Twitter Sentiment_zh	56.46	59.63
Biparty	51.08	51.26

Table 2 shows that the performance of different systems differs greatly on each of 14 target adjectives. For example, the accuracy of 大 da "large" is 95.53% by one system but only 46.51% by another system.

Table 2: The scores of 14 ambiguous adjectives

Words	Ins#	Max%	Min%	Stdev
大 |large	559	95.53	46.51	0.155
多|many	222	95.50	49.10	0.152
高 ||high	546	95.60	54.95	0.139
厚 |thick	20	95.00	35.00	0.160
深 |deep	45	100.00	51.11	0.176
重|heavy	259	96.91	34.75	0.184
巨大 |huge	49	100.00	10.20	0.273
重大 |great	28	100.00	7.14	0.243
小 |small	290	93.10	49.66	0.167
少 few	310	95.81	41.29	0.184
低 |low	521	93.67	48.37	0.147
薄 |thin	33	100.00	18.18	0.248
浅 |shallow	8	100.00	37.50	0.155
轻 |light	26	100.00	34.62	0.197

3.2 Systems

In this section, we give a brief description of the systems.

YSC-DSAA This system creates a new word library named SAAOL (SAA-Oriented Library), which is built manually with the help of software. SAAOL consists of positive words, negative words, NSSA, PSSA, and inverse words. The system divides the sentences into clauses using heuristic rules, and disambiguates SAA by analyzing the relationship between SAA and the keywords.

HITSZ_CITYU This group submitted three systems, including one baseline system and two improved systems.

HITSZ_CITYU_3: The baseline system is based on collocation of opinion words and their targets. For the given adjectives, their collocations are extracted from People's Daily Corpus. With human annotation, the system obtained 412 positive and 191 negative collocations, which are regarded as seed collocations. Using the context words of seed collocations as features, the system trains a one-class SVM classifier.

HITSZ_CITYU_2 and HITSZ_CITYU_1: Using HowNet-based word similarity as clue, the authors expand the seed collocations on both ambiguous adjectives side and collocated targets side. The authors then exploit sentence-level opinion analysis to further improve performance. The strategy is that if the neighboring sentences on both sides have the same polarity, the ambiguous adjective is assigned as the same polarity; if the neighboring sentences have conflicted polarity, the SO of ambiguous adjective is determined by its context words and the transitive probability of sentence polarity. The two systems use different parameters and combination strategy.

OpAL This system combines supervised methods with unsupervised ones. The authors employ Google translator to translate the task dataset from Chinese to English, since their system is working in English. The system explores three types of judgments. The first one trains a SVM classifier based on NTCIR data and EmotiBlog annotations. The second one uses search engine, issuing queries of "noun + SAA + AND + non-ambiguous adjective". The non-ambiguous adjectives include positive set ("positive, beautiful, good") and negative set ("negative, ugly, bad"). An example is "price high and good". The third one uses "too, very-

rules". The final result is determined by the majority vote of the three components.

CityUHK This group submitted four systems. Both machine learning method and lexicon-based method are employed in their systems. In the machine learning method, maximum entropy model is used to train a classifier based on the Chinese data from NTCIR opinion task. Clause-level and sentence-level classifiers are compared. In the lexicon-based method, the authors classify SAAs into two clusters: intensifiers (our positive-like adjectives in (4)) and suppressors (our negative-like adjectives in (5)), and then use the polarity of context to determine the SO of SAAs.

CityUHK4: clause-level machine learning + lexicon.

CityUHK3: sentence-level machine learning + lexicon.

CityUHK2: clause-level machine learning.

CityUHK2: sentence-level machine learning.

QLK_DSAA This group submitted two systems. The authors adopt their SELC model (Qiu, et al., 2009), which is proposed to exploit the complementarities between lexicon-based and corpus-based methods to improve the whole performance. They determine the sentence polarity by SELC model, and simply regard the sentence polarity as the polarity of SAA in the sentence.

QLK_DSAA_NR: Based on the result of SELC model, they inverse the SO of SAA when it is modified by negative terms. Our task includes only positive and negative categories, so they replace the neutral value obtained by SELC model by the predominant polarity of the adjective.

QLK_DSAA_R: Based on the result of QLK_DSAA_NR, they add a rule to cope with two modifiers 偏 pian "specially" and 太 tai "too", which always have the negative meaning.

Twitter sentiment This group submitted three systems. The authors use a training data collected from microblogging platform. By exploiting Twitter, they collected automatically a dataset consisting of negative and positive expressions. The sentiment classifier is trained using Naive Bayes with n-grams of words as features.

Twitter Sentiment: Translating the task dataset from Chinese to English using Google translator, and then based on training data in English texts from Twitter.

Twitter Sentiment_ext: With Twitter Sentiment as basis, using extended data.

Twitter Sentiment_zh: Based on training data in Chinese texts from Twitter.

Biparty This system transforms the problem of disambiguating SAAs to predict the polarity of target nouns. The system presents a bootstrapping method to automatically build the sentiment lexicon, by building a nouns-verbs biparty graph from a large corpus. Firstly they select a few nouns as seed words, and then they use a cross inducing method to expand more nouns and verbs into the lexicon. The strategy is based on a random walk model.

4 Discussion

The experimental results of some systems are promising. The micro accuracy of the best three systems is over 93%. Therefore, the inter-annotator agreement (91%) is not an upper bound on the accuracy that can be achieved. On the contrary, the experimental results of some systems are disappointing, which are below our predefined simple baseline (61.20%), and are only a little higher than random baseline (50%). The accuracy variance of different systems makes this task more interesting.

The participating 8 teams exploit totally different methods.

Human annotation. In YSC-DSAA system, the word library of SAAOL is verified by human. In HITSZ_CITYU systems, the seed collocations are annotated by human. The three systems rank top 3. Undoubtedly, human labor can help improve the performance in this task.

Training data. The OpAL system employs SVM machine learning based on NTCIR data and EmotiBlog annotations. The CityUHK systems trains a maximum entropy classifier based on the annotated Chinese data from NTCIR. The Twitter Sentiment systems use a training data automatically collected from Twitter. The results show that some of these supervised methods based on training data cannot rival unsupervised ones, partly due to the poor quality of the training data.

English resources. Our task is in Chinese. Some systems use English resources by translating Chinese into English, as OpAL and Twitter Sentiment. The OpAL system achieves a quite good result, making this method a promising direction. This also shows that disambiguating SAAs is a common problem in natural language.

5 Conclusion

This paper describes task 18 at SemEval-2010, disambiguating sentiment ambiguous adjectives. The experimental results of the 16 participating systems are promising, and the used approaches are quite novel.

We encourage further research into this issue, and integration of the disambiguation of sentiment ambiguous adjectives into applications of sentiment analysis.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 60703063), National Social Science Foundation of China (No. 08CYY016), and the Open Projects Program of Key Laboratory of Computational Linguistics(Peking University) , Ministry of Education. We thank Miaomiao Wen and Tao Guo for careful annotation of the data.

References

Ding X., Liu B. and Yu, P. 2008. A holistic lexicon-based approach to opinion mining. *Proceedings of WSDM'08*.

Hatzivassiloglou, V. and McKeown, K. 1997 Predicting the semantic orientation of adjectives. *Proceedings of ACL'97*.

Kim, S and Hovy, E. 2004. Determining the sentiment of opinions. *Proceedings of COLING'04*.

Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*.

Qiu L., Zhang W., Hu, C. and Zhao, K. 2009. SELC: A self-supervised model for sentiment analysis. In *Proceedings of CIKM'09*.

Takamura, H., Inui,T. and Okumura, M. 2006. Latent Variable Models for Semantic Orientations of phrases. *Proceedings of EACL'06*.

Takamura, H., Inui,T. and Okumura, M. 2007. Extracting Semantic Orientations of Phrases from Dictionary. *Proceedings of NAACL HLT '07*.

Turney, P. and Littman, M. 2003. Measuring praise and criticism: inference of semantic orientation from association. *ACM transaction on information systems*.

Wilson, T., Wiebe, J. and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of HLT/EMNLP'05*.