HITSZ_CITYU: Combine Collocation, Context Words and Neighboring Sentence Sentiment in Sentiment Adjectives Disambiguation

Ruifeng Xu[1,2], Jun Xu[1]
[1]Harbin Institute of Technology,
Shenzhen Campus, China
xuruifeng@hitsz.edu.cn
hit.xujun@gmail.com

Chunyu Kit[2]
[2]City University of Hong Kong,
Hong Kong
ctckit@cityu.edu.hk

Abstract

This paper presents the HIT_CITYU systems in Semeval-2 Task 18, namely, disambiguating sentiment ambiguous adjectives. The baseline system (HITSZ_CITYU_3) incorporates bi-gram and n-gram collocations of sentiment adjectives, and other context words as features in a one-class Support Vector Machine (SVM) classifier. To enhance the baseline system, collocation set expansion and characteristics learning based on word similarity and semi-supervised learning are investigated, respectively. The final system (HITSZ_CITYU_1/2) combines collocations, context words and neighboring sentence sentiment in a two-class SVM classifier to determine the polarity of sentiment adjectives. The final systems achieved 0.957 and 0.953 (ranked 1st and 2nd) macro accuracy, and 0.936 and 0.933 (ranked 2nd and 3rd) micro accuracy, respectively.

1 Introduction

Sentiment analysis is always puzzled by the context-dependent sentiment words that one word brings positive, neutral or negative meanings in different contexts. Hatzivassiloglou and Mckeown (1997) predicated the polarity of adjectives by using the pairs of adjectives linked by consecutive or negation conjunctions. Turney and Littman (2003) determined the polarity of sentiment words by estimating the point-wise mutual information between sentiment words and a set of seed words with strong polarity. Andreevskaia and Bergler (2006) used a Sentiment Tag Extraction Program to extract sentiment-bearing adjectives from WordNet. Esuli and Sebasian (2006) studied the context-dependent sentiment words in WordNet but ignored the instances in real context. Wu et al. (2008) applied collocation plus a SVM classifier in Chinese sentiment adjectives disambiguation. Xu et al. (2008) proposed a semi-supervised learning algorithm to learn new sentiment word and their context-dependent characteristics.

Semeval-2 Task 18 is designed to provide a common framework and dataset for evaluating the disambiguation techniques for Chinese sentiment adjectives. The HITSZ_CITYU group submitted three runs corresponding to one baseline system and one improved systems (two runs). The baseline system (HITSZ_CITYU_3) is based on collocations between sentiment words and their targets as well as their context words. For the ambiguous adjectives, 412 positive and 191 negative collocations are built from a 100-million-word corpus as the seed collocation set. Using the context words of seed collocations as features, a one-class SVM classifier is trained in the baseline system. Using HowNet-based word similarity as clue, the seed collocations are expanded to improve the coverage of collocation-based technique. Furthermore, a semi-supervised learning algorithm is developed to learn new collocations between sentiment words and their targets from raw corpus. Finally, the inner sentence features, such as collocations and context words, and the inter sentence features, i.e. neighboring sentence sentiments, are incorporated to determine the polarity of ambiguous adjectives. The improved systems (HITSZ_CITYU_1/2) achieved 0.957 and 0.953 macro accuracy (ranked 1st and 2nd) and 0.936 and 0.933 micro accuracy (ranked 2nd and 3rd), respectively. This result shows that collocation, context-words and neighboring sentence sentiment are effective in sentiment adjectives disambiguation.

The rest of this paper is organized as follows. Section 2 presents the collocation extraction subsystem based on lexical statistics. Section 3

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 448–451,
Uppsala, Sweden, 15-16 July 2010.

presents the baseline system and Section 4 presents the improved systems. The experiment results are given in Section 5 and finally, Section 6 concludes.

2 Collocation Extraction

A lexical statistics-based collocation extraction subsystem is developed to identify both the bi-gram and n-gram collocations of sentiment adjectives. This subsystem is based on our previous research on Chinese collocation extraction. It recognizes the co-occurring words of a headword as collocations which have co-occurrence frequency significance among all co-occurring words and co-occurrence position significance among all co-occurring positions.

For a sentiment adjective, noted as w_{head}, any word within the [-5,+5] context window is a co-word, denoted as $w_{co\text{-}i}$ for $1 \le i \le k$, where k is the total number of different co-words of w_{head}.

BI-Strength(w_{head},$w_{co\text{-}i}$) between a head word w_{head} and a co-word $w_{co\text{-}i}$ ($i=1$, *to* k) is designed to measure the co-occurrence frequency significance as follows:

$$BI-Strength(w_{head}, w_{co-i}) = 0.5 \cdot \frac{f(w_{head} w_{co-i}) - \overline{f(w_{head})}}{f_{\max}(w_{head}) - f_{\min}(w_{head})} + 0.5 \cdot \frac{f(w_{head} w_{co-i}) - \overline{f(w_{co-i})}}{f_{\max}(w_{co-i}) - f_{\min}(w_{co-i})} \quad (1)$$

where, $f_{max}(w_{head})$, $f_{min}(w_{head})$ and $\overline{f(w_{head})}$ are the highest, lowest and average co-occurrence frequencies among all the co-words of w_{head}, respectively; $f_{max}(w_{co\text{-}i})$, $f_{min}(w_{coi})$ and $\overline{f(w_{co-i})}$ are respectively the highest, lowest and average co-occurrence frequencies of the co-words for $w_{co\text{-}i}$. The value of *BI-Strength*(w_{head} $w_{co\text{-}j}$) ranges from -1 to 1, and a larger value means a stronger association. Suppose $f(w_{head}$,$w_{co\text{-}i}$, $m)$ is the frequency that $w_{co\text{-}i}$ co-occurs with w_{head} at position m($-5<=m<=5$). The *BI-Spread*(w_{head},$w_{co\text{-}i}$) is designed to characterizes the significance that $w_{co\text{-}i}$ around w_{head} at neighbouring places as follows:

$$BI-Spread(w_{head}, w_{co-i}) = \frac{\sum_{m=-5}^{5} | f(w_{head}, w_{co-i}, m) - \overline{f(w_{head}, w_{co-i})} |}{\sum_{m=-5}^{5} f(w_{head}, w_{co-i}, m)} \quad (2)$$

where, $\overline{f(w_{head}, w_{co-i})}$, $f_{max}(w_{head,}$,$w_{co\text{-}i})$, and f_{min} ($w_{hea,}$,$_{d}w_{co\text{-}i}$) are the average, highest, and lowest co-occurrence frequencies among all 10 positions, respectively. The value of *BI-Spread*(w_{head}, $w_{co\text{-}i}$) ranges from 0 to 1. A larger value means that w_{head} and $w_{co\text{-}i}$ tend to co-occur in one or two positions.

The word pairs satisfying, (1) *BI-Strength*(w_{head} $w_{co\text{-}j}$)>K_0 and (2) *BI-Spread*(w_{head}, $w_{co\text{-}i}$)>U0, are extracted as bi-gram collocations, where K_0 and U_0 are empirical threshold.

Based on the extracted bi-gram collocations, the appearance of each co-word in each position around w_{head} is analyzed. For each of the possible relative distances from w_{head}, only words occupying the position with a probability greater than a given threshold T are kept. Finally, the adjacent words satisfying the threshold requirement are combined as n-gram collocations.

3 The Baseline System

The baseline system incorporates collocation and context words as features in a one-class SVM classifier. It consists of two steps:

STEP 1: To match a test instance containing seed collocation set. If the instance cannot be matched by any collocations, go to **STEP 2**.

STEP 2: Use a trained classifier to indentify the sentiment of the word.

The collocations of 14 testing sentiment adjectives are extracted from a 100-million-word corpus. Collocations with obvious and consistent sentiment are manually identified. 412 positive and 191 negative collocations are established as the seed collocation set.

We think that the polarity of a word can be determined by exploiting the association of its co-occurring words in sentence. We assume that, the two instances of an ambiguous sentiment adjectives that have similar neighboring nouns may have the same polarity. Gamon and Aue (2005) made an assumption to label sentiment terms.

We extract 13,859 sentences containing collocations between negative adjective and targets in seed collocation set or collocations between ambiguous adjective and negative modifier (such as 过于 *too*) as the training data. These sentences are assume negative. A single-class classifier is then trained to recognize negative sentences. Three types of features are used:

(1) Context features include bag of words within context in window of [-5, +5]
(2) Collocation features contain bi-grams in window [-5,+5]
(3) Collocation features contain n-grams in window [-5,+5]

In our research, SVM with linear kernel is employed and the open source SVM package – LIBSVM is selected for the implementation.

4 The Improved System

The preliminary experiment shows that the baseline system is not satisfactory, especially the

coverage is low. It is observed that the seed collocation set covers 17.54% of sentences containing the ambiguous adjectives while the collocations between adjective and negative modifier covers only 11.28%. Therefore, we expand the sentiment adjective-target collocation set based on word similarity and a semi-supervised learning algorithm orderly. We then incorporate both inner-sentence features (collocations, context words, etc.) and inter-sentence features in the improved systems for sentiment adjectives disambiguation.

4.1 Collocation Set Expansion based on Word Similarity

First, we expand the seed collocation set on the target side. The words strongly similar to known targets are identified by using a word similarity calculation package, provided by HowNet (a Chinese thesaurus). Once these words co-occur with adjective within a context window more often than a threshold, they are appended to seed collocation set. For example, "低-技能(*low capacity*)" is expanded from a seed collocation "低-能力 (*low capacity*)".

Second, we manually identify the words having the same "trend" as the testing adjectives. For example, "上升 *increase*" is selected as a same-trend word of "高 *high*". The collocations of "上升" are extracted from corpus. Its collocated targets with confident and consistent sentiment are appended to the sentiment collocation set of "高" if they co-occurred with "高" more than a threshold. In this way, some low-frequency sentiment collocation can be obtained.

4.2 Semi-supervised Learning of Sentiment Collocations

A semi-supervised learning algorithm is developed to further expand the collocation seed set, which is described as follows. (It is revised based on our previous research (Xu et al. 2008). The basic assumption here is that, the sentiment of a sentence having ambiguous adjectives can be estimated based on the sentiment of its neighboring sentences.

Input: Raw training corpus, labeled as S_u,

Step 1. The sentences holding strong polarities are recognized from S_u which satisfies any two of following requirements, (1) contains known context-free sentiment word (CFSW); (2) contains more than three known context-dependent sentiment words (CDSW); (3) contains collocations between degree adverbs and known CDSWs; (4) contains collocations between degree adverbs and opinion operators (the verbs indicate a opinion operation, such as 称赞 *praise*); (5) contains known opinion indicator and known CDSWs.

Step 2. Identify the strong non-opinionated sentences in S_u. The sentences satisfying all of following four conditions are recognized as non-opinionated ones, (1) have no known sentiment words; (2) have no known opinion operators; (3) have no known degree adverbs and (4) have no known opinion indicators.

Step 3. Identify the opinion indicators in the rest sentences. Determine their polarities if possible and mark the conjunction (e.g.和 *and*) or negation relationship (e.g.但 but) in the sentences.

Step 4. Match the CFSWs and known CDSWs in S_u. The polarities of CFSWs are assigned based on sentiment lexicon.

Step 5. If a CDSW occurs in a sentence with certain orientations which is determined by the opinion indicators, its polarity is assigned as the value suggested. If a CDSW co-occur with a seed collocated target, it polarity is assigned according to the seed sentiment collocation set. Otherwise, if a CDSW co-occur with a CFSW in the same sentence, or the neighboring continual or compound sentence, the polarity of CDSW is assigned as the same as CFSW, or the reversed polarity if a negation indicator is detected.

Step 6. Update the polarity scores of CDSWs in the target set by using the cases where the polarity is determined in Step 5.

Step 7. Determine the polarities of CDSWs in the undetermined sentences. Suppose S_i is a sentence and the polarity scores of all its CFSWs and CDSWs are known, its polarity, labeled as $Plo(S_i)$, is estimated by using the polarity scores of all of the opinion words in this sentence, viz.:

$$Plo(Si) = \sum \begin{matrix} P\_\text{pos}(CFSW) - P\_neg(CFSW) \\ + P\_pos(CDSW) - P\_neg(CDSW) \end{matrix} \quad (3)$$

A large value (>0) of $Plo(s_i)$ implies that s_i tends to be positive, and vice versa.

Step 8. If the sentence polarity cannot be determined by its components, we use the polarity of its neighboring sentences s_{j-1} and s_{j+1}, labeled as $Plo(s_{j-1})$ and $Plo(s_{j+1})$, respectively, to help determine $Plo(s_j)$, viz.:

$$Plo(s_j) = 0.5 \cdot Plo(s_{j-1}) + Plo^*(s_j) + 0.5 \cdot Plo(s_{j+1}) \quad (4)$$

where, $Plo^*(s_j)$ is the polarity score of S_j (Following Equation 3) but ignore the contribution of testing adjectives while 0.5 are empirical weights.

Step 9. After all of the polarities of known CDSWs in the training data are determined, update the collocation set by identifying co-occurred pairs with consistent sentiment.
Step 10. Repeat Step 5 to Step 9 to re-estimate the sentiment of CDSWs and expand the collocation set, until the collocation set converge.

In this way, the seed collocation set is further expanded and their sentiment characteristics are obtained.

4.3 Sentiment Adjectives Classifier

We incorporate the following 8 groups of features in a linear-kernel two-class SVM classifier to classify the sentences with sentiment adjectives into positive or negative:

(1) The presence of known positive/negative opinion indicator and opinion operator
(2) The presence of known positive/negative CFSW
(3) The presence of known positive/negative CDSW(exclude the testing adjectives)
(4) The presence of known positive/negative adjective-target bi-gram collocations
(5) The presence of known positive/negative adjective-target n-gram collocations
(6) The coverage of context words surrounding the adjectives in the context words in training positive/negative sentences
(7) The sentiment of -1 sentence
(8) The sentiment of +1 sentence

The classifier is trained by using the sentences with determined sentiment which is obtained in the semi-supervised learning stage.

5 Evaluations and Conclusion

The ACL-SEMEVAL task 18 testing dataset contains 14 ambiguous adjectives and 2,917 instances. HITSZ_CITYU group submitted three runs. Run-1 and Run-2 are two runs corresponding to the improved system and Run-3 is the baseline system. The achieved performances are listed in Table 1.

Run ID	Marco Accuracy	Micro Accuracy
1	0.953	0.936
2	0.957	0.933
3(baseline)	0.629	0.665

Table 1: Performance of HITSZ_CITYU Runs

It is observed that the improved systems achieve promising results which is obviously higher than the baseline. They are ranked 1st and 2nd in Macro Accuracy evaluation and 2nd and 3rd in Micro Accuracy evaluation among 16 submitted runs, respectively.

6 Conclusion

In this paper, we proposed similarity-based and semi-supervised based methods to expand the adjective-target seed collocation set. Meanwhile, we incorporate both inner-sentence (collocations and context words) and inter-sentence features in a two-class SVM classifier for the disambiguation of sentiment adjectives. The achieved promising results show the effectiveness of collocation features, context words features and sentiment of neighboring sentences. Furthermore, we found that the neighboring sentence sentiments are important features for the disambiguation of sentiment ambiguous adjectives, which is obviously different from the traditional word sense disambiguation that emphasize the inner-sentence features.

References

Andreevskaia, A. and Bergler, S. 2006. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of EACL 2006*, pp. 209-216

Esuli, A. and Sebastian, F. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceeding of LREC 2006*, pp. 417-422.

Hatzivassiloglou, V. and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceeding of ACL 1997*, pp.174-181

Michael Gamon and Anthony Aue. 2005. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms. In *Proceedings of the ACL05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pp.57-64

Ruifeng Xu, Kam-Fai Wong et al. 2008. Learning Knowledge from Relevant Webpage for Opinion Analysis, in *Proceedings of 2008 IEEE / WIC / ACM Int. Conf. Web Intelligence*, pp. 307-313

Turney, P. D. and Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, vol. 21, no. 4, pp.315-346

Yunfang Wu, Miao Wang and Peng Jin. 2008. Disambiguating sentiment ambiguous adjectives, In *Proceedings of Int. Conf. on Natural Language Processing and Knowledge Engineering 2008*, pp. 1-8