

Building Domain-oriented Sentiment Lexicon by Improved Information Bottleneck

Weifu Du¹ and Songbo Tan²

¹Haerbin Institute of Technology, Haerbin, China

²Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

{duweifu,tansongbo}@software.ict.ac.cn

ABSTRACT

This paper describes an adapted information bottleneck approach for construction of domain-oriented sentiment lexicon. The basic idea is to use three kinds of relationships (WW_{inter} , WD_{inter} and WD_{intra}) to infer the semantic orientation of the out-of-domain words. The experimental results demonstrate that proposed method could dramatically improve the accuracy of the baseline approach on the construction of out-of-domain sentiment lexicon.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing

General Terms

Algorithms, Performance, Experimentation

Keywords

Sentiment Analysis; Opinion Mining; Information Retrieval

1. INTRODUCTION

In recent years, we have seen a rapid growth in non-topical text analysis, in which characterizations are sought of the opinions, feelings, and attitudes expressed in a text, rather than just the subjects. A key problem in this area is sentiment classification [11-15], where a document is labeled as a positive or negative evaluation of a target object (film, book, product, etc.).

In order to classify the sentiment of reviews in different domains, one of the commonly-used methods is to build a general sentiment lexicon. A particular challenge for construction of general sentiment lexicon is that the sentiment expression often behaves with strong domain-specific nature, and this so-called domain-specific nature makes it an important job to design an automated approach that could build a sentiment lexicon for each new domain.

So far, two kinds of approaches have been proposed to deal with this problem. One is based on a thesaurus. This method utilizes synonyms or glosses of a thesaurus to determine polarity of words [2][5][6][8]. The second approach exploits raw corpus. Polarity is decided by using co-occurrence in a corpus. This approach is

based on a hypothesis that polar terms conveying the same polarity tend to co-occur with each other. Typically, a small set of paradigm polar terms are prepared, and new polar terms are detected based on the strength of co-occurrence with the seeds [4][7][10].

Most of existing approaches take the homogeneous relationship between words (i.e., relationship between out-of-domain words and in-domain words (WW_{inter} -Relationship)) into account, while ignore the other two kinds of heterogeneous relationships (i.e., relationship between out-of-domain words and out-of-domain documents (WD_{intra} -Relationship), relationship between out-of-domain words and in-domain documents (WD_{inter} -Relationship)). Consequently, there is a lot of room for improvement and it is still a challenge to find more beneficial guidance from in-domain data for the construction of out-of-domain sentiment lexicon.

In this study, we take the construction of domain-oriented sentiment lexicon as clustering of sentiment words and extend the information-bottleneck clustering algorithm [9] by integrating more restriction for building an appropriate knowledge context of every sentiment word. First, some corpus-based techniques are employed to build three graphs (i.e., WW_{inter} -Graph, WD_{inter} -Graph and WD_{intra} -Graph) to reflect semantic relationships between sentiment words and documents. Second, an improved information bottleneck based clustering process is imposed on the three graphs, the polarity of out-of-domain sentiment words are identified simultaneously by the polar labels of in-domain data. The experimental results indicate that proposed approach can dramatically improve the performance of the baseline approach on the construction of out-of-domain sentiment lexicon.

2. PROPOSED ALGORITHM

The proposed domain-oriented sentiment lexicon construction method consists of two steps: (1) three graphs are built to reflect the semantic relationship between words and documents, either in-domain or out-of-domain; (2) based on the three graphs, an information bottleneck based clustering process is imposed to obtain the semantic orientation of every out-of-domain word.

2.1 Graph Building

2.1.1 Word-to-Word Graph (WW_{inter} -Graph)

Given the in-domain word collection $T_i = \{t_j \mid 1 \leq j \leq m\}$ and out-of-domain word collection $T_o = \{t_j \mid 1 \leq j \leq n\}$ of a document, the semantic similarity between any two words t_i and t_o can be

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'09, November 2-6, 2009, Hong Kong, China.

Copyright 2009 ACM 978-1-60558-512-3/09/11...\$10.00.

computed using approaches that are either knowledge-based or corpus-based.

In this study, we simply choose the mutual information to compute the semantic similarity between word t_i and t_o , as follows:

$$\text{sim}(t_i, t_o) = \log \frac{N \times p(t_i, t_o)}{p(t_i) \times p(t_o)} \quad (1)$$

which indicates the degree of statistical dependence between t_i and t_o . Here, N is the total number of words in the corpus and $p(t_i)$ and $p(t_o)$ are respectively the probabilities of the occurrences of t_i and t_o , i.e. $\text{count}(t_i)/N$ and $\text{count}(t_o)/N$, where $\text{count}(t_i)$ and $\text{count}(t_o)$ are the frequencies of t_i and t_o . $p(t_i, t_o)$ is the probability of the co-occurrence of t_i and t_o within a window with a predefined size k , i.e. $\text{count}(t_i, t_o)/N$, where $\text{count}(t_i, t_o)$ is the number of the times t_i and t_o co-occur within the window.

We use an adjacency matrix $A=[A_{ij}]_{m \times 2}$ to describe $\text{WW}_{\text{inter-Graph}}$, in which A_{il} denotes the total similarity between out-of-domain word t_o and all in-domain words with positive polar label; A_{l2} denotes the total similarity between out-of-domain word t_o and all in-domain words with negative polar label. Then A is normalized to \tilde{A} make the sum of each row equal to 1.

2.1.2 Word-to-Document Graph ($\text{WD}_{\text{inter-Graph}}$ and $\text{WD}_{\text{intra-Graph}}$)

Given the in-domain document collection $D = \{d_i | 1 \leq i \leq m\}$ and the out-of-domain word collection $T = \{t_j | 1 \leq j \leq n\}$ of an in-domain document, we can build a weighted bipartite graph G_{DW} from D and T in the following way: if word t_j appears in document d_i , we then create an edge between d_i and t_j . A nonnegative weight $\text{aff}(d_i, t_j)$ is specified on the edge, which is proportional to the importance of word t_j in document d_i , computed as follows:

$$\text{aff}(d_i, t_j) = \frac{tf_{t_j} \times idf_{t_j}}{\sum_{t \in d_i} tf_t \times idf_t} \quad (2)$$

where t represents a unique term in d_i and tf_t , idf_t are respectively the term frequency in the document and the inverse document frequency. We use an adjacency matrix $B=[B_{ij}]_{m \times n}$ to describe $\text{WD}_{\text{intra-Graph}}$ with each entry B_{ij} corresponding to $\text{aff}(d_i, t_j)$.

Similarly, B is normalized to \tilde{B} to make the sum of each row equal to 1.

Similar to the $\text{WW}_{\text{inter-Graph}}$, we can build an undirected graph $\text{WD}_{\text{inter-Graph}}$ to reflect the heterogeneous relationship between out-of-domain words and in-domain documents. We use an adjacency matrix $C=[C_{ij}]_{m \times 2}$ to describe $\text{WD}_{\text{inter-Graph}}$, in which

C_{i1} denotes the total similarity between out-of-domain word t_o and all in-domain documents with positive polar label; C_{i2} denotes the total similarity between out-of-domain word t_o and all in-domain documents with negative polar label. Then C is normalized to \tilde{C} to make the sum of each row equal to 1.

2.2 Sentiment Lexicon Construction

The information bottleneck algorithm (IB) is proposed by Slonim and Tishby [9], which is a clustering algorithm.

Input: normalized matrix \tilde{A} ($\text{WW}_{\text{inter-Graph}}$), \tilde{B} ($\text{WD}_{\text{intra-Graph}}$) and \tilde{C} ($\text{WD}_{\text{inter-Graph}}$).

Output: the semantic orientation of every out-of-domain sentiment words.

Initialization:

- Construct $\tilde{W}_o \equiv W_o$
- $\forall i, j = 1 \dots |\tilde{W}_o|, i < j$, calculate

$$d_{i,j} = \left(p(\tilde{w}_a) + p(\tilde{w}_b) \right) \left\{ D_{JS} \left[p(d_o | \tilde{w}_a), p(d_o | \tilde{w}_b) \right] + \alpha D_{JS} \left[p(d_i | \tilde{w}_a), p(d_i | \tilde{w}_b) \right] + \alpha D_{JS} \left[p(w_i | \tilde{w}_a), p(w_i | \tilde{w}_b) \right] \right\}$$

Loop:

- For $m = |\tilde{W}_o| - 1 \dots 1$
 - Find the indices $\{i, j\}$ for which $d_{i,j}$ is minimized
 - Merge $\{\tilde{w}_a, \tilde{w}_b\} \Rightarrow \tilde{w}_a$
 - Update $\tilde{W}_o = \{\tilde{W}_o - \{\tilde{w}_a, \tilde{w}_b\}\} \cup \{\tilde{w}_a\}$
 - Update $d_{i,j}$ costs w.r.t. \tilde{w}_a
- End For

Figure 1: Pseudo-code of the adapted information bottleneck method

The IB principle determines the distortion measure between the points x and c to be the Kullback-Leibler divergence between the conditional distributions $p(y|x)$ and $p(y|c)$,

$$D_{KL}[p(y|x) || p(y|c)] = \sum_y p(y|x) \log \frac{p(y|x)}{p(y|c)} \quad (3)$$

and use the Jensen-Shannon divergence to measure the merge cost.

$$\delta I(c_i, c_j) \equiv \left(p(c_i) + p(c_j) \right) \cdot D_{JS} \left[p(y|c_i), p(y|c_j) \right] \quad (4)$$

Where the functional D_{JS} is the Jensen-Shannon divergence defined as

$$D_{JS} \left[p_i, p_j \right] = \pi_i D_{KL} \left[p_i || \hat{p} \right] + \pi_j D_{KL} \left[p_j || \hat{p} \right] \quad (5)$$

where in our case

$$\left\{ \begin{aligned} \{p_i, p_j\} &\equiv \{p(y|c_i), p(y|c_j)\} \\ \{\pi_i, \pi_j\} &\equiv \left\{ \frac{p(c_i)}{p(c_*)}, \frac{p(c_j)}{p(c_*)} \right\} \\ \hat{p} &= \pi_i p(y|c_i) + \pi_j p(y|c_j) \end{aligned} \right. \quad (6)$$

We extend the traditional IB algorithm for more in-domain knowledge (sentiment polarity label), and the pseudo-code of the adapted information bottleneck algorithm is shown in Figure 1. This algorithm iteratively searches a clustering for the out-of-domain words, and assigns sentiment polarity labels to the word clusters to complete the sentiment-lexicon building task.

3. EXPERIMENTAL SETUP

In order to evaluate the properties of the proposed algorithm, in this section, we describe our experiments and the data used in these experiments. Aimed at Chinese applications, we conduct the experiments based on the specialty of Chinese language, and verify the performance on Chinese web reviews. However, the main proposed approach in this paper is language independent in essence.

3.1 Data

We download texts from the Internet, which including comments on hotel (from www.ctrip.com), electronics (from <http://detail.zol.com.cn/>) and stock (from <http://blog.sohu.com/stock/>). The detailed information is illustrated in Table 1,

Table 1 the detail information of corpus

Domain	Positive	Negative	Total
Hotel	2000	2000	4000
Electronics	1054	554	1608
Stock	364	683	1047

We use ICTCLAS (<http://ictclas.org/>), a Chinese word segmentation software, to extract sentiment words from these texts. In the usage of the part-of-speech tagging function provided by this software, we take all adjectives, adverbs and adjective-noun phrases as candidate sentiment words.

Table 2 the detail information of labeled sentiment lexicon of each domain

Extracted Sentiment Words	Total (before pruning)	Non-Repeated			
		Pos		Neg	
		Independ	Depend	Independ	Depend
Hotel	93616	253	93	199	59
Electronics	58967	298	124	242	90
Stock	79560	343	89	567	112

After removing the repeated words and words with ambiguity, we get a list of words in each domain. Then, we manually label the semantic orientation of every word, and use these labeled word lists as the sentiment lexicons in the following experiments. In order to highlight the nature of domain sentiment lexicon, we distinguish the domain-dependent sentiment words from the domain-independent sentiment words in the process of labeling. We take the words only occur in one domain or the ones show reverse orientation among different domains as domain-dependent sentiment words. To justify the reliability of this labeling process,

we ask three annotators to label one domain data, respectively. Three annotators had pair-wise agreement scores[1] of 80.10%, 83.87% and 85.96%, which is high enough to be considered consistent. Table 2 presents the detailed information of labeled sentiment lexicon of each domain.

3.2 Comparison Method

Since proposed method aims to construct domain-specific sentiment lexicon, we should compare it with existing word semantic orientation inferring methods. Most of these approaches infer word semantic orientation by measuring the relationship between words, which can be either corpus-based or knowledge-based, Since the proposed approach is also corpus-based, for justness, we take the PMI method[10], improved PMI (SM+SO) method[3] and lexicon extension (LE) method[7] as the baseline methods, and compare the performance between these methods and our method.

The PMI method takes some labeled sentiment words as paradigm words to infer the semantic orientation of unlabelled words. In the implementation, we use the common part (sentiment words) of in-domain data and out-of-domain data as the paradigm words of the PMI method.

For SM+SO method and LE method, we set up the experimental environment as the default configurations as Gamon and Aue [3] and Kanayama and Nasukawa [7].

3.3 Evaluation Metrics

We use accuracy to evaluate the performance of proposed method. Let C be the clustering function which maps from word (or document) to its true sentiment label, and F be the function which maps from word to its prediction sentiment label that given by the sentiment inferring methods. The accuracy is defined as:

$$Accuracy(w) = \frac{|\{w | w \in W_o \wedge C(w) = F(w)\}|}{|W_o|} \quad (7)$$

4. EXPERIMENTAL RESULTS

4.1 Performance Comparison

Table 3 and Table 4 report the performance comparison between proposed method and the three baselines on six tasks for domain-independent words and domain-dependent words.

By the comparison between the two tables, we can find that nearly all approaches show better performance on domain-independent tasks than on domain-dependent tasks, which indicates the difficulty of domain sentiment lexicon construction.

From Table 4, we can find that proposed method shows better performance on nearly all of the data sets. In consideration of that the baseline methods take only the relationship between out-of-domain words and in-domain words (WW_{inter} -Relationship) into account, while neglect the other two kinds of relationship (WD_{inter} -Relationship and WD_{intra} -Relationship), the full use of the three kinds of relationship may contributes to the performance of proposed method.

Seen from these experimental results, a question may arise: why does the PMI method perform so poorly that it seems to disaccord the conclusion drawn by Turney and Littman[10].

A reasonable explanation is that the PMI method is corpus-based, and the corpus size influences its performance very much. The experimental results provided by Turney and Littman [10] is

obtained by making use of search engine, and taking the whole Internet as the corpus. While the corpus in our experiment is relatively small, which may brings much noise and makes the co-occurrence information sparse. These factors may lead to the poor performance of the PMI method. From another perspective, this also shows the robustness of proposed method on relatively small-scaled corpus.

Table 3 Accuracy of domain-independent sentiment word classification

	Baselines			Proposed Method
	PMI	SM+SO	LE	
Elec→Htl	76.6	77.5	80.7	88.1
Elec→Sto	69.7	68.3	71.3	73.6
Htl→Elec	74.1	76.7	83.4	79.7
Htl→Sto	85.4	88.0	86.7	84.8
Sto→Elec	70.5	73.3	81.3	76.7
Sto→Htl	67.9	71.2	81.8	84.8
Average	74.4	75.4	80.8	81.2

Table 4 Accuracy of domain-dependent sentiment word classification

	Baselines			Proposed Method
	PMI	SM+SO	LE	
Elec→Htl	68.4	73.5	73.2	87.5
Elec→Sto	57.8	60.6	63.1	73.2
Htl→Elec	72.1	75.4	76.3	75.9
Htl→Sto	73.7	76.4	78.1	82.2
Sto→Elec	70.6	73.3	73.4	74.1
Sto→Htl	68.8	71.2	73.6	82.8
Average	68.5	71.7	72.9	79.2

4.2 Varying the Parameter

There is only one parameter in proposed method, which is the trade-off parameter α in Equation 11. We conduct experimental tests by varying the parameter on the three data sets: electronics-to-hotel, hotel-to-electronics and stock-to-hotel. Figure 2 presents this experimental result.

The parameter α reflects the influence of in-domain knowledge on the guide of clustering on out-of-domain data. From this figure, we can find that when α is small, by introducing in-domain knowledge, the accuracy increase; while when α is larger than a threshold, the algorithm gradually degenerates into the clustering based fully on in-domain knowledge, which excludes the contextual knowledge about out-of-domain data (i.e., WD_{intra} -Relationship). This will result in the decline in accuracy. According to this figure, we set α to 0.25 in our experiments.

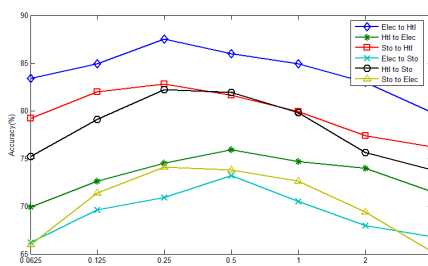


Figure 2 Accuracy curves of word classification vs. α

5. CONCLUSIONS

In this paper, we propose an adapted information bottleneck method for the automatic construction of domain-oriented sentiment lexicon by fusing the cross-domain knowledge and within-domain knowledge in a unified information-theoretic framework, and solve this problem using an iterative reinforcement approach. In our experiment, it is shown that proposed method greatly outperforms the baseline methods in the task of building out-of-domain sentiment lexicon.

6. ACKNOWLEDGMENTS

This work was mainly supported by two funds, i.e., 0704021000 and 60803085.

7. REFERENCES

- [1] J. Cohen. A coefficient of agreement for nominal scales. In Educational and Psychological measurements, 1960, 37-46
- [2] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. CIKM 2005
- [3] M. Gamon and A. Aue. Automatic identification of sentiment vocabulary exploiting low association with known sentiment terms. ACL 2005
- [4] V. Hatzivassiloglou and K. McKeown. Predicting the semantic orientation of adjectives. ACL.1997
- [5] M. Hu and B. Liu. Mining and summarizing customer reviews. KDD 2004
- [6] J. Kamps, M. Marx, R. Mokken, etc. Using WordNet to measure semantic orientation of adjectives. LREC 2004
- [7] H. Kanayama and T. Nasukawa. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. EMNLP 2006.
- [8] S. Kim and E. Hovy. Determining the sentiment of opinions. COLING 2004.
- [9] N. Slonim and N. Tishby. Agglomerative information bottleneck. NIPS 1999.
- [10] P. Turney and M. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. In: ACM Transactions on Information Systems. 2003.
- [11] H. Tang, S. Tan and X. Cheng. A Survey on Sentiment Detection of Reviews. Expert Systems with Applications. 2009.
- [12] S. Tan, G. Wu, H. Tang and X. Cheng. A novel scheme for domain-transfer problem in the context of sentiment analysis. CIKM 2007.
- [13] S. Tan, X. Cheng, Y. Wang, H. Xu. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. ECIR 2009.
- [14] Q. Wu, S. Tan, H. Zhai, G. Zhang, M. Duan and X. Cheng. SentiRank: Cross-Domain Graph Ranking for Sentiment Classification. WI 2009.
- [15] A. Kennedy and D. Inkpen. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. Computational Intelligence. 2006.