Integrating Intra- and Inter-document Evidences for Improving Sentence Sentiment Classification

ZHAO Yan-Yan[1] QIN Bing[1] LIU Ting[1]

Abstract Sentence sentiment classification is an important task of sentiment analysis. It aims to classify the sentences into positive, negative, or objective. One can consider sentence sentiment classification as a standard text categorization problem. However, determining the sentiment orientation of a review sentence requires more than the features inside the sentence itself, especially for the sentences with little or ambiguous inside sentence features. Through observing, some features outside the sentence can interact with its inside features to enhance the overall performance of sentence sentiment classification. Thus in this paper, we propose two such outside sentence features: intra-document evidence and inter-document evidence. Then in order to improve the sentence sentiment classification performance, a graph-based propagation approach is presented to incorporate these inside and outside sentence features. The experimental results on camera domain show that the proposed approach performs better than the approaches without using outside sentence features, and outperforms other representational previous approaches.

Key words Sentence sentiment classification, sentiment analysis, intra-document evidence, inter-document evidence, graph-based propagation approach

DOI 10.3724/SP.J.1004.2010.01417

Sentence sentiment classification, which aims to classify the review sentences into positive, negative, or objective, has received considerable attention recently. Generally speaking, this task can be implemented using either supervised or unsupervised methods. To date, most of these methods were performed by applying a standard classification algorithm on each sentence in isolation. For example, in supervised methods, many features inside a sentence are used to train a classifier, such as n-gram lexical features, syntactic features, and so on[1−3]; and in unsupervised methods, the sentiment features such as polarity words inside a sentence are used to yield a score to determine the sentence′s final sentiment orientation[4−6]. However, determining the sentiment orientation of a review sentence requires more than the features inside the sentence, especially for the sentences with ambiguous inside sentence features. Two examples are listed as follows.

Exampe 1. The sentence "*富士相机还加入了10倍光学变焦*" ("*Another plus on Fuji′s side is the 10X optical zoom*" in English) reflects positive sentiment orientation. But it is hard to determine its sentiment orientation, since it contains ambiguous features (such as context independent polarity words) inside this sentence.

Exampe 2. We present three sentences with the context dependent polarity word "*长*" ("*long*" in English). The first sentence "*电池寿命很长*" ("*The battery life is very long*" in English) reflects positive orientation; and in contrast, the second sentence "*这个相机的启动时间很长*" ("*The camera has long startup*" in English) reflects negative orientation; and the third sentence "*昨天，我沿着一条很长的路逛街*" ("*Yesterday, I went shopping along a long street*" in English) reflects objective orientation. Since the features (such as the polarity word "*long*") inside are ambiguous, it is hard to determine their accurate sentiment orientations.

The above examples indicate that in order to identify a review sentence′s sentiment orientation, just focusing on the features inside a sentence is far from enough. Features outside a sentence need to be explored, so as to interact with its inside features to enhance the performance. In this paper, two kinds of features outside a sentence are explored. The first is the intra-document evidence. For a target sentence, it refers to its surrounding sentences′ sentiment orientations in the same document, such as the sentences before or after the target sentence. Experimentally speaking, this kind of feature is inspired by the observation that text spans occurring near to each other may share the same orientation status[3−4].

The second kind of feature is the inter-document evidence. It refers to the orientations of the sentences that are semantically similar to the target sentence in other topical-related documents. This kind of feature is inspired by the observation that the sentences with similar semantics may have the same sentiment orientation, even if they appear in different documents. For example, the two sentences "*色彩还原真实*" ("*The color rendition is true*" in English) and "*色彩表现非常真实*" ("*Color performance is very true*" in English) are semantically similar. Thus, they share the same orientation status "positive".

To our knowledge, the ideas resembling to the intra-document evidence have already been proposed. In supervised methods, it is often represented as feature vectors to be incorporated in classifiers (such as conditional random fields (CRFs))[3]. Also, in unsupervised methods, the sentences with ambiguous features can make reference to their context′s orientations[4]. But to date, there is little work that exploits the proposed inter-document evidence. Moreover, there is little work to model these outside and inside sentence features to interact with each other for sentence sentiment classification.

In order to better integrate these outside and inside sentence features, a new graph-based approach is proposed. In this approach, a sentence graph is created with all the test sentences being nodes, and by connecting two sentence nodes if the intra- or inter-document evidence exists. After multiple iterative computations through a propagation algorithm Potts model[7], each node in the graph is finally annotated with one of the three orientation states (positive, negative, and objective). Finally, we apply the proposed graph-based approach to a collection of Chinese review sentences in camera domain. Experimental results demonstrate that it outperforms the approaches without using the outside sentence features, and meanwhile outperforms other representational previous approaches.

The remainder of this paper is organized as follows. Section 1 introduces the related work. Section 2 describes the

Manuscript received September 8, 2009; accepted May 20, 2010
Supported by National High Technology Research and Development Program of China (863 Program) (2008AA01Z144) and National Natural Science Foundation of China (60803093, 60975055)
1. Research Center for Information Retrieval, Harbin Institute of Technology, Harbin 150001, P. R. China

proposed graph-based propagation approach, which integrates intra- and inter-document evidences. Sections 3 and 4 present the experiments and results. Finally, we conclude this paper in Section 5.

1 Related work

1.1 Sentence sentiment classification

Most of the previous related researches have focused more on the features inside a sentence, either using supervised methods or unsupervised methods. Supervised methods consider the determining of sentence sentiment orientation as a classification problem and solve it using machine learning models and various features. Gamon[1] experimented with a range of different feature sets for customer feedback data, such as surface features and linguistic features. Kim et al.[2] adopted three types of features, including lexical features, positional features, and polarity word features. Unsupervised methods mainly use the sentiment features inside a sentence, such as the polarity words, to yield a score to determine the sentiment orientation for each sentence. Yu et al.[6] first determined the semantically oriented words in the sentence, then used the average per word log-likelihood scores to measure the sentence semantic orientation. Similarly, Kim et al.[5] found the polarity words using WordNet, WSJ Data and Columbia Wordlist with a strength ranging from -1 to $+1$, and added all of them to compute a positive/negative score for the sentence. Since these supervised/unsupervised methods deeply rely on the inside sentence features, it is hard to identify the sentiment orientations of the sentences with ambiguous inside features such as Examples 1 and 2.

Our use of the intra-document evidence for sentence sentiment classification is preceded by plenty of existing research work. Zhao et al.[3] used CRF model to integrate this kind of feature into feature vectors. For the sentences with ambiguous inside sentence features, Hu et al.[4] considered their nearby sentences′ orientations. However, to our knowledge, there is little work which integrates inter-document feature in the form of graph for sentence sentiment classification. For other tasks in sentiment analysis, there are also some researches using the intra-document evidence. Pang et al.[8] proposed a graph-based model, which used min-cut algorithms based on nearby sentences to help smooth the decision surface for subjectivity detection. McDonald et al.[9] reduced the joint sentence and document level analysis to a sequential classification problem using constrained Viterbi inference. Because of the usefulness of the intra-document evidence, we also adopt it as an important outside sentence feature. However, there is little work that uses inter-document evidence.

1.2 Potts model

If a variable can have more than two values and there is no ordering relation between the values, the graph comprised of nodes denoting such variables can be solved by Potts model[7]. Potts model can be used for an arbitrary number of classes, and it has been widely used in several applications such as image restoration[10] and rumor transmission[11].

Recently, Potts model is successfully used in sentiment analysis tasks. Takamura et al.[12] used Potts model for classifying semantic orientations of phrases (pairs of an adjective and a noun). They constructed a lexical graph by connecting similar words, and adopted the Potts model for the probability model of the lexical work. The experimental results demonstrated the effectiveness of this model.

Compared with other graph based propagation algorithm[13−15], Potts model is more suitable for describing and modeling the graph constructed by intra- and inter-document evidences in our paper.

2 The proposed approach

2.1 Overview

We divide the sentences into two categories, namely, unambiguous sentences and ambiguous sentences, according to whether their inside sentence features are unambiguous or not. For the unambiguous sentences (such as the sentence "*佳能相机拍的照片相当不错*" ("*The image of Canon camera is very great*" in English)), we can easily identify their sentiment orientations through their unambiguous inside sentence features, such as the polarity word "*不错*" ("*great*" in English) that carries "positive" orientation. However, for the ambiguous sentences (such as the sentence "*富士相机还加入了10倍光学变焦*" ("*Another plus on Fuji′s side is the 10X optical zoom*" in English)), it is difficult to identify the sentiment orientation only with its inside sentence features. But fortunately, this kind of sentence is often related to some unambiguous sentences. Such as its neighboring sentences in the same document or other semantically similar sentences in other topically related documents, whose inside sentence features are perhaps unambiguous. Therefore, the sentiment orientation of an ambiguous sentence can be determined with the help of these so called outside sentence features. In this paper, two outside sentence features, namely, intra- and inter-document evidences are proposed to help to identify these sentences′ sentiment orientations.

As mentioned above, the sentiment orientation of a target sentence is commonly related to its nearby text spans that are in the same document (intra-document evidence). In particular, in this paper, this kind of feature refers to the orientations of the two sentences that are before and after the target sentence. Besides that, we can also consider the sentences in other documents that are semantically similar to the target sentence (inter-document evidence). They can also be used to help to judge the sentiment orientation of the target sentence, especially when the documents they belong to are about the same or similar topics.

For example, Table 1 illustrates samples of the intra- and inter-document evidences for a target sentence "*色彩还原真实*" ("*The color rendition is true*" in English) in camera domain. Intra-document evidences are indicated by the left column, in which $Intra_{Bf}$ and $Intra_{Af}$ represent the sentences before and after the target sentence in the same document. Obviously, the target sentence′s sentiment orientation is the same as $Intra_{Af}$. On the other hand, inter-document evidences are indicated by the right column. As can be seen, five sentences $Inter_1 \sim Inter_5$ from different documents are regarded as the inter-document evidences for the target sentence, all of which share the same sentiment orientation "positive". Although these semantically similar sentences are from different documents, they can also help to identify the target sentence′s sentiment orientation.

In order to make full use of the unambiguous sentences, and then to better integrate these two outside sentence features, a graph-based propagation algorithm, Potts model, is adopted. This graph is constructed with all the test sentences being nodes and by connecting two nodes if the intra- or inter-document evidence exists between them. Fig. 1 shows the structure of the sentence graph. The nodes in the structure represent all the test sentences. Some of the nodes are considered as seed nodes, which refer to the unambiguous sentences. Then, these seed nodes combined with the intra- and inter-document evidences (represented as the links) to determine the sentiment orientations of the ambiguous sentences.

Table 1 Samples of the intra- and inter-document evidences for a target sentence "色彩还原真实" ("*The color rendition is true*" in English.)

Intra-document evidence	Inter-document evidence
$Intra_{Bf}$. 使用 第二 代	$Inter_1$. 还原 真实
DIGI 图像 引擎	(The rendition is true.)
(It uses the second	$Inter_2$. 色彩 表现 很 真实
generation of DIGI	(Color performance is very true.)
Image Engine.)	$Inter_3$. C70 的 肤色 还原 较 真实
$Intra_{Af}$. 更 适合 东方人	(The skin color rendition of C70 is true.)
(It is more suitable for	$Inter_4$. 色彩 还原 准确
Asians.)	(The color rendition is accurate.)
	$Inter_5$. 另外 Z1 的 色彩 还原 我 觉得 也 可以
	(I think the color rendition of Z1 is ok.)

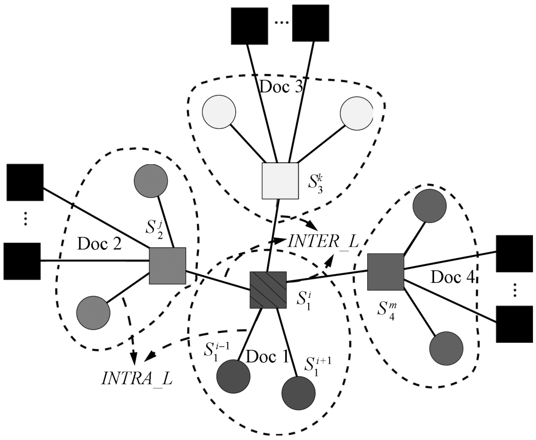


Fig. 1 The structure of sentence graph

Each link between two nodes in the sentence graph, is either intra-document evidence link $INTRA\_L$ or inter-document evidence link $INTER\_L$, as shown in Fig. 1. For the target sentence S_1^i (the i-th sentence in Doc 1), sentence S_1^{i-1} and S_1^{i+1} in Doc 1 are its intra-document evidences, the links between them are regarded as $INTRA_L$. The sentence S_2^j in Doc 2, S_3^k in Doc 3, and S_4^m in Doc 4 are the inter-document evidences for the target sentence S_1^i, the links between them are regarded as $INTER_L$. The sentiment orientation of the target sentence S_1^i is then partly determined by the sentiment orientations of all its relevant sentences (such as S_1^{i-1}, S_1^{i+1}, S_2^j, S_3^k, and S_4^m), besides its own inside sentence features.

Every node in the sentence graph has three optional states (positive, negative, and objective). Potts model is then adopted to model and compute the probability of each state for every unseeded node, with the help of the unseeded node′s related nodes and links. Finally, the state with the highest probability is considered as the node′s sentiment orientation in one iteration. After multiple iterations, we can acquire a stable state for each node.

In short, the framework of our graph-based approach mainly contains two steps. The first step is the graph initialization, which consists of the seed node identification and the link weight setting. The second step is to use Potts model to estimate the state for each sentence in the graph, so as to fulfill the sentence sentiment classification task.

2.2 Graph initialization

2.2.1 Seed node identification

Seed nodes are needed for the sentence graph. Since the unambiguous sentences can be easily and exactly identified by their inside features, we consider them as the seed nodes.

In this paper, we adopt a syntactic path-based method to find the seed nodes. Through analyzing the sentiment features inside a sentence, we find that the collocation of the polarity word and its corresponding target can be considered as unambiguous features for determining a sentence′s sentiment orientation. We call the collocation "P_T_Pair". As shown in Fig. 2, the polarity word in isolation does not always show sentiment orientation (such as "好" ("*good*" in English)). But the polarity word "差" ("*bad*" in English) in the P_T_Pair"差_质量" ("*bad_quality*" in English), can clearly determine the sentence sentiment orientation "negative", according to its own sentiment orientation. Therefore we can automatically find the seed nodes by recognizing and analyzing P_T_Pairs in sentences.

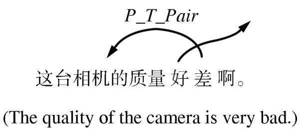
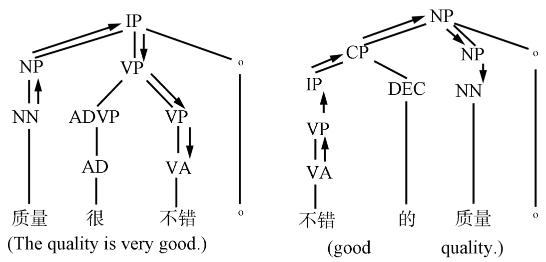


Fig. 2 An example containing P_T_Pair

Through analysis, we investigate that the path in a parse tree connecting a polarity word and its corresponding target can better describe the relationship between them. We use this kind of syntactic path to find P_T_Pairs. Figs. 3 (a) and 3 (b) respectively show the syntactic paths in two parse trees. In the left sentence, the syntactic path between target "质量" ("*quality*" in English) and polarity word "不错" ("*good*" in English) is "subject-copula-predicate" relationship, which can be represented as NN↑NP↑IP↓VP↓VP↓VA. In the right sentence, the syntactic path between polarity word "不错" ("*good*" in English) and target "质量" ("*quality*" in English) is "attribute-head" relation, which can be represented as VA↑VP↑IP↑CP↑NP↓NP↓NN.



(a) Syntactic path in Sen 1 (b) Syntactic path in Sen 2

Fig. 3 Examples of syntactic paths

From Fig. 3, we can find that syntactic paths can better describe the relationships between polarity words and their targets, which can be considered as the evidences to acquire P_T_Pairs. Meanwhile, the sentiment orientations ("positive") for the two sentences can be obtained from the orientations of their own P_T_Pairs (e.g., "不错_质量" ("*good_quality*" in English)).

We give the detailed process for identifying seed nodes as follows.

Step 1. Get frequent syntactic paths.

We collect 600 sentences from web consisting of three domains. The polarity words of these sentences are labeled through an existing sentiment lexicon, which includes 1 478 commonly used polarity words. And then the corresponding targets are labeled manually. Afterwards, we use Dan Bikel′s phrase parser to parse these sentences, and acquire all the syntactic paths between the polarity words and their corresponding targets. We analyze the frequency of each kind of syntactic path, and select 33 top frequent syntactic paths as final rules, according to a threshold 3.

Step 2. Get positive and negative seed nodes.

For a given sentence after syntactic parsing, all of the P_T_Pairs can be acquired through matching with the syntactic path rules. If the polarity word of P_T_Pair appears in the existing sentiment lexicon, the sentiment orientation of the polarity word is used as the sentence′s final sentiment orientation. In addition, we also consider whether there is a negation word such as "不", "不是" ("*not*" or "*no*" in English), appearing closely around the polarity word (window size is 5). If so, the final sentence sentiment orientation is inverted. And for the sentences containing more than two conflicting P_T_Pairs, we abandon them as seed nodes. Then we can get all the positive and negative seed nodes.

Step 3. Get objective seed nodes.

If a sentence satisfies all the following conditions, we experientially consider it as an objective seed node.

C 1. The sentence has intra-document evidences, and does not have any inter-document evidence.

C 2. The sentence does not have any P_T_Pair.

C 3. The sentence does not have polarity words.

C 4. The sentence does not have negation words.

Through the above three steps, we can get all the three kinds of seed nodes. In other words, we complete the identification for the unambiguous sentences.

Then, we shall initialize the probabilities for each node in the graph. The three state (positive, negative, and objective) probabilities for each seed node are set according to its given orientation. For example, if the given orientation is "positive", we set the probabilities to 0.8, 0.1, 0.1 for positive, negative, and objective state, respectively. For the left unseeded sentences, we can simply set the probability to 1/3 for each state without using any inside sentence features. Besides, we can also selectively set the probability $Pr_{state}^{classifier}(s_i)$ for each state considering the sentence s_i′s inside features, where $Pr_{state}^{classifier}(s_i)$ denotes the classifier′s (such as maximum entropy, or support vector machine (SVM), and so on) estimate of the probability when sentence s_i carries state.

2.2.2 Link weight setting

After identifying seed nodes, the sentence sentiment classification task can be made to only focus on the identification of ambiguous sentences. Then the seed nodes combined with the intra- and inter-document evidences are used for identifying the orientations of these sentences. Correspondingly, two kinds of links, $INTRA\_L$ and $INTER\_L$, are proposed to convey the intra- and inter-document evidences, respectively. For $INTRA\_L$, we just consider the sentiment orientations of the sentences before and after the target sentence, and set them with the weight w_{intra}. We also consider the transitional words such as "但是" ("*although*" in English). If there are some transitional words in the $INTRA\_L$ related sentences or the target sentence, we set the link weights between them to $-w_{intra}$.

On the other hand, for $INTER\_L$, it refers to the link between semantically similar sentences. Thus two kinds of strategies are proposed as follows.

Strategy 1. If the two sentences share the same P_T_Pair, they can be considered semantically similar to each other.

As mentioned above, the orientation for a sentiment sentence is mainly conveyed by the P_T_Pairs. Therefore, if two sentences share the same P_T_Pairs, they always show the same orientation. Such as the two sentences in Fig. 3, both of them contain the same *P_T_Pair*, and accordingly, they share the same sentiment orientation "positive". So if two sentences contain the same P_T_Pairs, we link them in the graph and set this kind of link weight to w_{inter_1}. Besides, we also consider the negation words, such as "没有", "不" ("*no*" or "*not*" in English). If one sentence contains the negation word and the other does not, we set the link weight between them to $-w_{inter\_1}$.

Strategy 2. A similarity computing method is applied to compute the similarity between two sentences. If the similarity is more than a predefined threshold th, they can be considered semantically similar to each other.

This strategy is an easy and direct way to judge whether two sentences are semantically similar. There are many kinds of methods to compute sentence similarities. In this paper, we choose a simple but effective method: cosine similarity method. Given two vectors of attributes, A and B, the cosine similarity is represented as:

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{1}$$

For our task, we split the sentence into words to construct a word vector. Therefore in (1), A represents the target sentence′s word vector and B represents the word vector of the sentence being compared. As the weight for each word vector is 1 or 0, then $A \cdot B$ refers to the number of shared words between word vector A and B. And, $\|A\|$ and $\|B\|$ refer to the number of words of A and B. Note that the stop words have been removed from the vector A and B before the similarity computation.

Based on these, the cosine similarity between a target sentence and another sentence can be computed. If the similarity exceeds the predefined threshold th, the corresponding sentence can be considered as its semantically similar sentence. Then we link them in the graph, and set the similarity to w_{inter_2} as the link weight.

2.3 Potts model for sentence sentiment classification

The sentence graph is created with all the test sentences being nodes, and by connecting two sentences if the intra- or inter-document evidence exists. We adopt Potts model for the probability model of the graph.

Suppose a graph consisting of nodes and weighted links is given. In this graph, the state class of a node i is represented by c; the weight between nodes i and j is represented by w_{ij}; and the probability of the state class c for node i is represented by $p_i(c)$. Then for each iteration, we first compute the three probabilities $p_i(c)$ for each node i; and then estimate the status of the whole graph through integrating all the nodes′ probabilities and the link weights between them; and finally judge whether the status of the whole graph tends to be stable. If not, go on the next iteration; otherwise, for each node, the state with the highest probability is considered as its final sentiment orientation.

The state probability $p_i(c)$ of node i, is then estimated recursively with the Potts model algorithm. It assigns $p_i(c)$ to each node according to the number of nodes connected to it as well as the strength of their connections. The equation calculating $p_i(c)$ of node i is shown as follows:

$$p_i(c) = \frac{\exp(\alpha \cdot \delta(c, a_i) + \beta \cdot \sum\limits_j \omega_{ij} \cdot p_j(c))}{\sum\limits_n \exp(\alpha \cdot \delta(n, a_i) + \beta \cdot \sum\limits_j \omega_{ij} \cdot p_j(n))} \tag{2}$$

where $p_j(c)$ ($j = 1, 2, \cdots, t$, $j \neq i$) are the probabilities of the nodes linking to node i; n represents three kinds of sentiment states; α is a positive constant representing a weight on seed nodes, and β is a constant called the inverse-temperature; a_i is the state of node i before update, and function δ returns 1 if two arguments c and a_i are equal to each other (i.e., this demonstrates that the state of node i does not change before and after update), and returns 0 otherwise. However, if node i is not a seed node, the function $p_i(c)$ can be obtained by removing $\alpha \cdot \delta(c, a_i)$ from (2). When node i is a seed node, α will be set to a very large value to reserve its original sentiment orientation.

After estimating the three probabilities for each node in the sentence graph, we use the variational free energy function $F(c)$ in (3) of Potts model to evaluate the status of the whole graph.

$$\begin{aligned} F(c) = & -\alpha \cdot \sum_i \sum_{c_i} p_i(c_i) \cdot \delta(c_i, a_i) - \\ & \beta \cdot \sum_{ij} \sum_{c_i, c_j} p_i(c_i) \cdot p_j(c_j) \cdot \omega_{ij} \cdot \delta(c_i, c_j) - \\ & \sum_i \sum_{c_i} -p_i(c_i) \cdot \log p_i(c_i) \end{aligned} \tag{3}$$

Potts model is an iterative calculation procedure. Therefore, the stable status for the whole sentence graph would be acquired through multiple iterations. Suppose when the difference in the value $F(c)$ for the whole graph is below a threshold (we set it to 1) before and after updating, the computation of this model converges. Thus, after multiple iterations until the whole graph converges, each node can obtain the final probabilities for the three states respectively, and the state with the highest probability is considered as the node′s final sentiment orientation.

3 Experimental setup

3.1 Data set and evaluation metrics

In order to assess the performance of the proposed approach, we adopt a sentence set including 5 617 sentences from a Chinese Opinion Analysis Evaluation (COAE)[1]. These sentences are acquired from 138 topically related Chinese reviews that are all about the camera domain. In this paper, we consider them as the nodes of a sentence graph and apply our graph-based method to determine the sentiment orientation for each sentence node. We randomly select 911 sentences as the development set and 949 sentences as the test set. Then two annotators are asked to manually label the sentiment orientation for each sentence. Cohen′s kappa, a measure of inter-annotator agreement ranging from 0 to 1, is 0.73, indicating a good strength of agreement[16].

As for the evaluation metrics, we adopt accuracy to evaluate the performances of sentence sentiment classification. It refers to the proportion of correctly identified sentences in all sentences. On the other hand, the precision P, recall R, and F-*score* are also computed in a standard manner to evaluate the performances for the positive, negative or objective sentences in the test sentence set, respectively.

3.2 Parameter estimation

In our approach, two kinds of link weights including three parameters (w_{intra_1}, w_{inter_1}, and th), need to be estimated. As all these parameters are used for evaluating the semantical similarity between two neighboring sentence nodes, we assume that their values are all within the range from 0 to 1. Then we tune up these parameters by using our graph-based method on the development data, and finally set w_{intra} to 1, set w_{inter_1} to 0.9 (selected from {0.1, 0.2, 0.3, ···, 1.0}), and set the similarity threshold th of w_{inter_2} to 0.4 (selected from {0.1, 0.2, 0.3, ···, 1.0}).

In order to apply the Potts model, we need to further estimate the parameter α and β mentioned in (2) and (3). Based on our graph-based method, we assume that the orientation labels of the seed sentences are reliable. Therefore, the parameter α is empirically set to 1 000, which can be large enough to hold the seed nodes′ initial orientations. And the other parameter β is set to 0.1 according to the experiments on the development data.

3.3 Comparison methods

We design three kinds of strategies for comparison.

Baseline. After we get all the three kinds of seed sentences based on the sentiment features inside a sentence, all the other sentences are annotated randomly.

Voting. After we get all the seeds, we can obtain the intra- and inter-document evidences for each node as its neighbors. Then we use a simple voting algorithm to get the final tag for each node. That is to say, a node is classified by a majority vote of its neighbors and assigned the tag that is most common amongst them. However, there also exist a few nodes that have no seed neighbors, we then randomly annotate them. In this way, if we only adopt intra- or inter-document evidences as neighbors, the method is presented as ***V*_*intra*** or ***V*_*inter***. And if we adopt both of them, this method is presented as ***V*_*i* + *i***.

Potts. After we get all the seed nodes, we can obtain the intra- and inter-document evidences for each node as its neighbors. If we just consider intra- or inter-document evidences, and acquire all the nodes′ sentiment orientations though Potts model, the method is presented as ***P*_*intra*** or ***P*_*inter***. And if we consider both of them, the method is turned into the graph-based method proposed in this paper, which can also be represented as ***P*_*i*+*i***.

In order to compare our method with other previous methods, we reimplement Hu and Liu′s method[4] and Kim and Hovy′s method[2].

Unsupervised. For each sentence, this method first acquires all its inside polarity words through matching the words in the same sentiment lexicon (Subsection 2.2.1). Then suppose that the one with "positive" is set to the value "+1", and the one with "negative" is set to "−1". The sum of all the existing polarity words′ values yields a

[1]COAE is an authentic evaluation in China, which includes many tasks. 22 related Chinese colleges and research institutes participated in this evaluation in 2008.

final score indicating this sentence′s sentiment orientation. If a sentence has no polarity words or the final score is zero, they simply consider its context sentences′ orientations.

Supervised. This method considers three kinds of inside sentence features including lexical features, positional features, and polarity word features. Then two classifiers, namely, a maximum entropy (ME)-based classifier and a support vector machine (SVM)-based classifier are respectively used to classify the sentences in the test data.

4 Results

4.1 Comparison with three strategies

Table 2 presents the performances of all the three above strategies and our approach (presented as $\boldsymbol{P}\_i+i$), which shows that our approach significantly outperforms every alternative strategy (*Z-test* with $P < 0.05$).

We discuss the comparative results in Table 2 as follows:

1) $\mathbf{V}\_intra$ and $\mathbf{P}\_intra$ are both adding the intra-document evidences besides the inside sentence features. Seen from Table 2, both of the strategies perform better than the *Baseline*, which is just based on the inside sentence features. This proves that the outside sentence feature: intra-document evidence is effective for improving sentence sentiment classification. Moreover, this feature breaks down the sentence boundary for sentence sentiment classification, and meanwhile validates the importance of the context.

2) $\mathbf{V}\_inter$ and $\mathbf{P}\_inter$ are both adding the inter-document evidences besides the inside sentence features. Seen from Table 2, both of the strategies perform better than the *Baseline*. It demonstrates that the inter-document evidence is effective. Besides, the proposed inter-document evidence first breaks down the document boundaries for sentence sentiment classification, and brings in new evidences from a set of topically related documents.

3) $\mathbf{V}\_i+i$ and $\mathbf{P}\_i+i$ both integrate the intra- and inter-document evidences into a joint model. Seen from Table 2, although the results of strategies with "+inter" are lower than strategies with "+intra" accordingly, their composite strategies get a significant performance. This can suggest that: a) intra- and inter-document evidences are complementary to each other, and b) integrating the two kinds of outside sentence features is advisable. Besides, we can also find that the features outside the sentence can interact with its inside features to enhance the final performance.

4) Comparing with the Voting related strategies, Potts model can better integrate the intra- and inter-document evidences. As the fourth and fifth rows in Table 2, $\mathbf{P}\_intra$, $\mathbf{P}\_inter$, and $\mathbf{P}\_i+i$ perform better than $\mathbf{V}\_intra$, $\mathbf{V}\_inter$, and $\boldsymbol{V}\_i+i$, respectively. This illustrates that Potts model is suitable for processing the sentence graph created in our graph-based method.

4.2 Comparison with previous work

In this section, we compare the performances of our graph-based method with the representational unsupervised and supervised methods. Table 3 shows the comparison results. The last column presents the accuracy of sentence sentiment classification on all the sentences for each kind of method. Besides, the performances for the three kinds (positive, negative, and objective) of sentences are also described by P, R, and *F-score*, respectively in other columns. Among them, the results for supervised method are gathered using 5-fold cross validation with one fold in 949 sentences (the test set mentioned in Section 3.1) for testing and the other 4 folds for training.

From the results shown in Table 3, we can obtain the following two conclusions.

1) Observing the accuracy for each kind of method, we can find that our graph-based method performs better than both of the unsupervised (*Z-test* with $P < 0.01$) and supervised method (*Z-test* with $P < 0.1$). It proves that our method, which integrates the inside and outside sentence features in the form of sentence graph, can better classify the sentence sentiment orientation, comparing with some previous work. During the experiments, we find that this unsupervised method strongly depends on the polarity words within a sentence. However, it is hard to find the correct polarity words in a sentence. Besides, since many positive or negative sentences include some ambiguous polarity words (such as "长"("*long*" in English)), it can confuse these sentences with the actual objective sentences. As for the supervised method, it considers other inside sentence features besides the polarity words, and then integrates them into a sentence classifier. However, the ambiguous sentences are also hard to identify, whether using the ME-based classifier or the SVM-based classifier. Furthermore, the outside sentence features are difficult to be integrated into the classification model, especially for the inter-document evidence.

Table 2 The comparison of the three kinds of strategies and our approach (P_i+i)

Methods	Accuracy (%)			
	+random	+intra	+inter	+intra+inter
Baseline	45.31 (*Baseline*)	–	–	–
Voting	–	56.27 ($\boldsymbol{V}\_intra$)	51.42 ($\boldsymbol{V}\_inter$)	62.28 ($\boldsymbol{V}\_i+i$)
Potts model	–	58.69 ($\boldsymbol{P}\_intra$)	55.85 ($\boldsymbol{P}\_inter$)	**67.23** ($\boldsymbol{P}\_i+i$)

Table 3 The comparative performances of the previous methods and our graph-based method

Methods	Positive sentences (%)			Negative sentences (%)			Objective sentences (%)			All sentences (%)
	P	R	*F-score*	P	R	*F-score*	P	R	*F-score*	Accuracy
Unsupervised	79.43	52.73	63.38	77.45	27.72	40.83	32.02	90.43	47.29	52.69
Supervised (SVM)	60.55	87.40	71.54	63.39	40.70	49.57	51.90	21.81	30.71	60.38
Supervised (ME)	68.06	72.06	70.00	59.06	57.19	58.11	41.42	37.23	39.22	60.70
Graph-based (our)	73.63	82.14	77.66	67.28	64.21	65.71	43.84	34.04	38.32	67.23
Ex-Graph based	73.18	82.56	77.59	68.15	64.56	66.31	45.07	34.04	38.79	67.54

2) Observing P, R, and F-*score* for each kind of method, we can find that for the positive or negative sentences, our graph-based method can yield good results, and perform better (F-*score*) than either of the unsupervised or supervised method by a wide margin. However, the performance for objective sentences is poor for every kind of these three methods. This is enough to illustrate that the objective sentences are too difficult to identify. The first reason is that the proportion of the objective sentences in all the review sentences is low and they are always confused with other positive/negative sentences. Thus sometimes, their intra-document evidences are inaccurate. Another reason is that the objective sentences are always expressed in a more free manner. Therefore, their inter-document evidences do not work well.

During the graph initialization procedure in our graph-based method, we simply set the unseeded nodes with the probability 1/3 for each state (positive, negative, or objective). However, this lacks the consideration of their inside sentence features. In order to integrate more informatively the initial probabilities for these unseeded sentences, we use the probabilities $Pr_{state}^{classifier}(s_i)$ for each state of every sentence s_i. $Pr_{state}^{classifier}(s_i)$ is computed by an ME classifier that is trained on the features in supervised method (shown in Section 3.3). The last row of Table 3 shows the results of this extended graph-based method (Ex-Graph based). However, the final accuracy of this Ex-Graph based method is just a little higher (0.31 % increased) than that of the graph-based method. This indirectly indicates that the initial probabilities of the unseeded nodes acquired from the classifier are not very effective, i.e., the supervised method is weak in processing these sentences with ambiguous inside sentence features.

4.3 Impact of seed nodes

As previously observed, seed nodes are important for sentence graph initialization. Generally speaking, there are two requirements that the seed nodes need to satisfy.

Requirement 1. In the sentence graph, the more accurate the seed nodes are, the better the performance is.

Requirement 2. In the sentence graph, the more the seed nodes are, the better the performance is.

We will testify whether the selected seed nodes in our method satisfy the above two requirements.

Table 4 shows the accuracies for all the three kinds of seed nodes based on our automatic seed node acquisition approach. The high accuracies (all above 90 %) suggest that *P_T_Pairs* are effective inside sentence features for recognizing seed nodes. Furthermore, the syntactic paths are effective for identifying the *P_T_Pairs* in sentences. And obviously, it also demonstrates that the sentences acquired by our approach are accurate enough to be considered as seed nodes.

Table 4 Accuracy for each kind of seed sentences

Seed category	Accuracy	#
Positive	91.64 %	813
Negative	94.84 %	410
Objective	90.31 %	298

Table 5 shows the quantity distributions of three kinds of sentences (positive, negative, and objective) in seed node set, development set, and test set, respectively. We can find that the quantity of seed nodes in the sentence graph is quite large, which contains 1 521 sentences and accounts for nearly 30 % of all sentences (5 617). This mainly benefits from our automatic seed nodes acquisition method. As far as we know, most related work usually selected the seeds manually[5, 12, 17], therefore, they were unable to get large quantity of seeds. From Table 5, we can also find that the quantity distribution (percentage) of these three kinds of sentences in seed node set almost has the same trend with that in the development set and test set. It demonstrates that our seed node distribution is reliable.

Table 5 The quantity distributions of three kinds of sentences in seed set, development set, and test set

Category	# of positive	# of negative	# of objective	# of sum
Seed set	813	410	298	1 521
Development set	461	243	207	911
Test set	476	285	188	949

In order to investigate the impact of the seed node quantity on sentence sentiment classification, we randomly select 20 %, 40 %, 60 %, and 80 % of the 1 521 existing seed nodes as seed nodes in the sentence graph to run the proposed approach $\boldsymbol{P}\_i+i$ (also can be represented by graph-based), respectively. Fig. 4 shows the performance curve changing with ratio R. We can find that the larger the quantity of the seed nodes is, the higher the accuracy is.

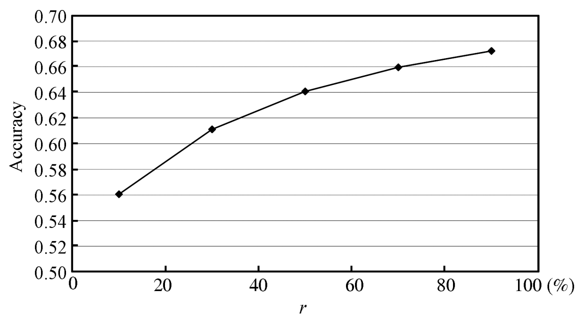


Fig. 4 The performance curve changing with R

4.4 Accuracy of intra- and inter-document evidences

In order to evaluate the accuracies of both intra- and inter-document evidences, we randomly select 300 intra-document evidences and 300 inter-document evidences from the sentence graph, and manually examine them. Table 6 shows the accuracy for each kind of outside sentence feature.

Table 6 Accuracy for each kind of outside sentence feature

Feature category	Accuracy	#
Intra-document evidence	77.33 %	300
Inter-document evidence	66.67 %	300

We can find that the intra-document evidence is shown with a higher accuracy. This can explain why so much previous work uses this kind of outside feature. Besides, the inter-document evidence is shown with a 66.67 % accuracy. Although it has been proved that the inter-document evidence is helpful for sentence sentiment classification, its accuracy is relatively not very high.

Through analyzing, the reason is found that the two strategies of extracting inter-document evidences in Section 2.2.2 have some strict limitations. For Strategy 1, when a sentence contains more than two P_T_Pairs, sometimes just one of P_T_Pairs determines its sentence sentiment orientation, while the others are useless. For instance, sentence Sen 1 "*该相机的照相效果不错*" ("*The camera′s picture is perfect*" in English) and sentence Sen 2 "*仅仅照相效果不错的相机并不一定是好相机*" ("*A camera only with perfect picture is not always a good camera*" in English) share the same P_T_Pair "*不错_照相效果*" ("*perfect_picture*" in English), but it is useless for determining the Sen 2′s sentiment orientation. Thus, the orientations of these two sentences are semantically different.

For Strategy 2, we just simply consider the words in a sentence as vectors, but sometimes, the deeper semantic knowledge needs to be considered. For example, the sentence Sen 3 "*色彩 还原 真实*" ("*The color rendition is true*" in English) and the sentence Sen 4 "*色彩 还原 失真*" ("*The color rendition is untrue*" in English) are similar according to cosine similarity algorithm, since only one word vector represented by "*真实*" ("*true*" in English) in Sen 3 and "*失真*" ("*untrue*" in English) in Sen 4 is different. However, the semantics of the two words are entirely opposite, which causes the orientations of Sen 3 and Sen 4 being opposite.

Based on these, more strict limitations and more semantic knowledge will be considered during the linkage construction procedure in future.

5 Conclusion and future work

We propose a new graph-based propagation approach for sentence sentiment classification. Different from previous work, our approach breaks down the sentence and even the document boundaries for this task, and brings in two new outside sentence features, namely, intra- and inter-document evidences. The contributions of this paper are as follows: 1) Intra- and inter-document evidences are proved effective on improving sentence sentiment classification; 2) Potts model is shown as suitable to integrate the inside and outside sentence features.

Besides, the framework designed in this paper (which includes three different models, namely, the seed node identification model, the linkage identification and setting model, and the graph-based propagation model), can be treated as a new plug-in system, i.e., all the three models are replaceable plug-ins for the sentence sentiment classification task. Thus, each model can be easily replaced by a more effective model in future.

Our future work will be carried out along the following directions:

1) From Table 6, we can find that the accuracies for outside sentence features are not very high, although the experimental results have demonstrated their effectiveness on sentence sentiment classification. Therefore in future work, more strict limitations and more semantic knowledge will be considered while selecting inter-document evidences. Further, more useful outside sentence features will be explored.

2) From Table 3, we can find that the performances for objective sentences are poor. The main reason is that objective sentences are always expressed in a free manner and without any obvious evidences. Thus they are hard to be distinguished from the positive/negative sentences. But through observing, most objective sentences are always found in the front position of a review. Therefore, in future, we will study the locational distributions of the objective sentences in the reviews, and moreover, study some methods to integrate these locational features into the sentence graph.

3) The graph-based method proposed in this paper is domain adaptive, so in future, we plan to extend more experiments for other domains besides camera domain.

References

1 Gamon M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland: Association for Computational Linguistics, 2004. 841−847

2 Kim S M, Hovy E. Automatic identification of pro and con reasons in online reviews. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. Sydney, Australia: Association for Computational Linguistics, 2006. 483−490

3 Zhao J, Liu K, Wang G. Adding redundant features for CRFs-based sentence sentiment classification. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Hawaii, USA: Association for Computational Linguistics, 2008. 117−126

4 Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA: ACM, 2004. 168−177

5 Kim S M, Hovy E. Automatic detection of opinion bearing words and sentences. In: Proceedings of the 2nd International Joint Conference on Natural Language Processing. Jeju Island, Korea: Springer, 2005. 61−66

6 Yu H, Hatzivassiloglou V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Sapporo, Japan: Association for Computational Linguistics, 2003. 129−136

7 Wu F Y. The potts model. *Reviews of Modern Physics*, 1982, **54**(1): 235−268

8 Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 2004. 271−278

9 McDonald R, Hannan K, Neylon T, Wells M, Reynar J. Structured models for fine-to-coarse sentiment analysis. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association for Computational Linguistics, 2007. 432−439

10 Tanaka K, Morita T. Application of cluster variation method to image restoration problem. In: Proceedings of the Theory and Applications of the Cluster Variation and Path Probability Methods. New York, USA: Association for Computational Linguistics, 1996. 353−373

11 Liu Z Z, Luo J, Shao C G. Potts model for exaggeration of a simple rumor transmitted by recreant rumormongers. *Physical Review E*, 2001, **64**(4): 34−43

12 Takamura H, Inui T, Okumura M. Extracting semantic orientations of phrases from dictionary. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. New York, USA: Association for Computational Linguistics, 2007. 292−299

13 Su F Z, Markert K. Subjectivity recognition on word senses via semi-supervised mincuts. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Boulder, Colorado: Association for Computational Linguistics, 2009. 1−9

14 Rao D L, Ravichandran D. Semi-supervised polarity lexicon induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece: Association for Computational Linguistics, 2009. 675−682

15 Lerman K, Blair-Goldensohn S, McDonald R. Sentiment summarization: evaluating and learning user preferences. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece: Association for Computational Linguistics, 2009. 514−522

16 Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, **20**(1): 37−46

17 Hatzivassiloglou V, McKeown K R. Predicting the semantic orientation of adjectives. In: Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics. Madrid, Spain: Association for Computational Linguistics, 1997. 174−181

ZHAO Yan-Yan Ph. D. candidate in the Department of Computer Science and Technology, Harbin Institute of Technology. Her research interest covers sentiment analysis. Corresponding author of this paper. E-mail: yyzhao@ir.hit.edu.cn

QIN Bing Professor in the Department of Computer Science and Technology, Harbin Institute of Technology. Her research interest covers text mining and sentiment analysis.
E-mail: bqin@ir.hit.edu.cn

LIU Ting Professor in the Department of Computer Science and Technology, Harbin Institute of Technology. His research interest covers natural language processing and information retrieval.
E-mail: tliu@ir.hit.edu.cn