Collocation Polarity Disambiguation Using Web-based Pseudo Contexts

Yanyan Zhao, Bing Qin and Ting Liu[*]
Harbin Institute of Technology, Harbin, China
`{yyzhao, bqin, tliu}@ir.hit.edu.cn`

Abstract

This paper focuses on the task of collocation polarity disambiguation. The *collocation* refers to a binary tuple of a polarity word and a target (such as ⟨long, battery life⟩ or ⟨long, startup⟩), in which the sentiment orientation of the polarity word ("long") changes along with different targets ("battery life" or "startup"). To disambiguate a collocation's polarity, previous work always turned to investigate the polarities of its surrounding contexts, and then assigned the majority polarity to the collocation. However, these contexts are limited, thus the resulting polarity is insufficient to be reliable. We therefore propose an unsupervised three-component framework to expand some pseudo contexts from web, to help disambiguate a collocation's polarity. Without using any additional labeled data, experiments show that our method is effective.

1 Introduction

In recent years, more attention has been paid to sentiment analysis as it has been widely used in various natural language processing applications, such as question answering (Wiebe et al., 2003; Yu and Hatzivassiloglou, 2003), information extraction (Riloff et al., 2005) and opinion-oriented summarization (Hu and Liu, 2004; Liu et al., 2005). Meanwhile, it also brings us lots of interesting and challenging research topics, such as subjectivity analysis (Riloff and Wiebe, 2003), sentiment classification (Pang et al., 2002; Kim and Hovy, 2005;

---

[*]Correspondence author: tliu@ir.hit.edu.cn

Wilson et al., 2009; He et al., 2011), opinion retrieval (Zhang et al., 2007; Zhang and Ye, 2008; Li et al., 2010) and so on.

One fundamental task for sentiment analysis is to determine the semantic orientations of words. For example, the word "beautiful" is positive, while "ugly" is negative. Many researchers have developed several algorithms for this purpose and generated large *static* lexicons of words marked with prior polarities (Hatzivassiloglou and McKeown, 1997; Turney et al., 2003; Esuli, 2008; Mohammad et al., 2009; Velikovich et al., 2010). However, there exist some polarity-ambiguous words, which can *dynamically* reflect different polarities along with different contexts. A typical polarity-ambiguous word "长" ("long" in English) is shown with two example sentences as follows.

1. 该相机的[电池寿命]^{t}很[长]^p。(Positive)

   Translated as: The [battery life]^t of this camera is [long]^p. (Positive)

2. 该相机的[启动时间]^{t}很[长]^p。(Negative)

   Translated as: This camera has [long]^p [startup]^t. (Negative)

The phrases marked with p superscript are the polarity-ambiguous words, and the phrases marked with t superscript are targets modified by the polarity words. In the above two sentences, the sentiment orientation of the polarity word "长" ("long" in English) changes along with different targets. When modifying the target "电池寿命" ("battery life" in English), its polarity is positive; and when modifying "启动时间" ("startup" in English), its polarity is

160

negative. In this paper, *we especially **define** the collocation as a binary tuple of the polarity-ambiguous word and its modified target*, such as ⟨长,电池寿命⟩ (⟨long, battery life⟩ in English) or ⟨长,启动时间⟩ (⟨long, startup⟩ in English). This paper concentrates on the task of **collocation polarity disambiguation**.

This is an important task as the problem of polarity-ambiguity is frequent. We analyze 4,861 common binary tuples of polarity words and their modified targets from 478 reviews[1], and find that over 20% of them are the collocations defined in this paper. Therefore, the task of collocation polarity disambiguation is worthy of study.

For a sentence s containing such a collocation c, since the in-sentence features are always ambiguous, it is difficult to disambiguate the polarity of c by using them. Thus some previous work turned to investigate its surrounding contexts' polarities (such as the sentences before or after s), and then assigned the majority polarity to the collocation c (Hatzivassiloglou and McKeown, 1997; Hu and Liu, 2004; Kanayama and Nasukawa, 2006). However, since the amount of contexts from the original review is very limited, the final resulting polarity for the collocation c is insufficient to be reliable.

Fortunately, most collocations may appear multiple times, in different forms, both within the same review and within topically-related reviews. Thus for a collocation, we can collect large amounts of contexts from other reviews to improve its polarity disambiguation. These expanded contexts are called ***pseudo contexts*** in this paper. Some previous work used the similar methods. For example, Ding (Ding et al., 2008) expanded some pseudo contexts from a topically-related review set. But since the review set is limited, the expanded contexts are still limited and unreliable. In order to overcome this problem, we propose an unsupervised three-component framework to expand more pseudo contexts from web for the collocation polarity disambiguation.

Without using any labeled data, experiments on a Chinese data set from four product domains show that the three-component framework is feasible and the web-based pseudo contexts are useful for the collocation polarity disambiguation. Compared to other previous work, our method achieves an *F1*

---

[1]The dataset will be introduced in Section 4.1 in detail.

score of 72.02%, which is about 15% higher.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 shows the proposed approach including three independent components. Section 4 and 5 presents the experiments and results. Finally we conclude this paper in Section 6.

2 Related Work

The key of the collocation polarity disambiguation task is to recognize the polarity word's sentiment orientation of a collocation. There are basically two types of approaches for word polarity recognition: corpus-based and dictionary-based approaches. Corpus-based approaches find co-occurrence patterns of words in the large corpora to determine the word sentiments, such as the work in (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Riloff and Wiebe, 2003; Turney et al., 2003; Kaji and Kitsuregawa, 2007; Velikovich et al., 2010). On the other hand, dictionary-based approaches use synonyms and antonyms in WordNet to determine word sentiments based on a set of seed polarity words. Such approaches are studied in (Kim and Hovy, 2006; Esuli and Sebastiani, 2005; Kamps et al., 2004). Overall, most of the above approaches aim to generate a large *static* polarity word lexicon marked with prior polarities.

However, it is not sensible to predict a word's sentiment orientation without considering its context. In fact, even in the same domain, a word may indicate different polarities depending on what targets it is applied to, especially for the polarity-ambiguous words, such as "长" ("long" in English) shown in Section 1. Based on these, we need to consider both the polarity words and their modified targets, i.e., the collocations mentioned in this paper, rather than only the polarity words.

To date, the task in this paper is similar with much previous work. Some researchers exploited the features of the sentences containing collocations to help disambiguate the polarity of the polarity-ambiguous word. For example, Hatzivassiloglou (Hatzivassiloglou and McKeown, 1997) and Kanayama (Kanayama and Nasukawa, 2006) used conjunction rules to solve this problem from large domain corpora. Suzuki (Suzuki et al., 2006)

161

took into account many contextual information of the word within the sentence, such as exclamation words, emoticons and so on. However, the experimental results show that these in-sentence features are not rich enough.

Instead of considering the current sentence alone, some researchers exploited external information and evidences in other sentences or other reviews to infer the collocation's polarity. For a collocation, Hu (Hu and Liu, 2004) analyzed its surrounding sentences' polarities to disambiguate its polarity. Ding (Ding et al., 2008) proposed a holistic lexicon-based approach of using global information to solve this problem. However, the contexts or evidences from these two methods are limited and unreliable. Except for the above unsupervised methods, some researchers (Wilson et al., 2005; Wilson et al., 2009) proposed supervised methods for this task, which need large annotated corpora.

In addition, many related works tried to learn word polarity in a specific domain, but ignored the problem that even the same word in the same domain may indicate different polarities (Jijkoun et al., 2010; Bollegala et al., 2011). And some work (Lu et al., 2011) combined difference sources of information, especially the lexicons and heuristic rules for this task, but ignored the important role of the context. Besides, there exists some research focusing on word sense subjectivity disambiguation, which aims to classify a word sense into subjective or objective (Wiebe and Mihalcea, 2006; Su and Markert, 2009). Obviously, this task is different from ours.

3 The Proposed Approach

3.1 Overview

The motivation of our approach is to make full use of web sources to collect more useful pseudo contexts for a collocation, whose original contexts are limited or unreliable. The framework of our approach is illustrated in Figure 1.

In order to disambiguate a collocation's polarity, three components are carried out:

1. Query Expansion and Pseudo Context Acquisition: This paper uses the collocation as query. For a collocation, three heuristic query expansion strategies are used to generate more flexible queries, which have the same or completely opposite polar-
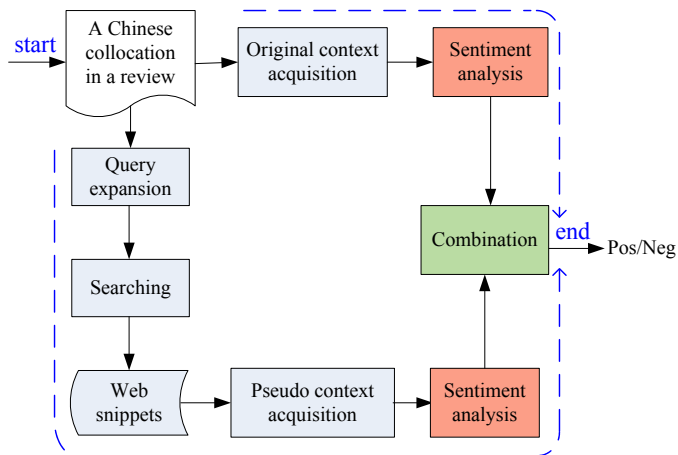

Figure 1: The framework of our approach.

ity with this collocation. Searching these queries in the domain-related websites, lots of snippets can be acquired. Then we can extract the pseudo contexts from these snippets.

2. Sentiment Analysis: For both original contexts and the expanded pseudo contexts from web, a simple lexicon-based sentiment computing method is used to recognize each context's polarity.

3. Combination: Two strategies are designed to integrate the polarities of the original and pseudo contexts, under the assumption that these two kinds of contexts can be complementary to each other.

It is worth noting that this three-component framework is flexible and we can try to design different strategies for each component. Next sections will give a simple example strategy for each component to show its feasibility and effectiveness.

3.2 Query Expansion and Pseudo Context Acquisition

3.2.1 Why Expanding Queries

For a collocation, such as ⟨长,电池寿命⟩ (⟨long, battery life⟩ in English), the most intuitive query used for searching is constructed by the form of "target + polarity word", i.e., 电池寿命长 (battery life long in English). Even if we search this query alone, a great many web snippets covering the polarity word and target will be retrieved. But why do we still need to expand the queries?

In fact, for a collocation, though the amount of the retrieved snippets is large, lots of them cannot provide accurate pseudo contexts. The reason is that the

162

polarity words in some snippets do not really modify the targets, such as in the sentence "The *battery life* is short, and finds few buyers for a *long* time." There exist no modifying relation between "*battery life*" and "*long*".

In order to filter these meaningless snippets, we can simply search with a new query "电池寿命长" by ***surrounding it with quotes*** (noted as **Strategy0**). However, this can drastically decline the amount of snippets. In addition, as the new query is short, in many retrieved snippets, there also exist no modifying relations between the polarity words and targets. As a result, if we just use this query strategy, the expanded pseudo contexts are limited and cannot yield ideal performance.

Therefore, we need to design some effective query expansion strategies to ensure that (1) the polarity words do modify the targets in the retrieved web snippets, and (2) the snippets are more enough.

3.2.2 Query Expansion Strategy

We first investigate the modifying relations between polarity words and the targets, and then construct effective queries.

Observed from previous work (Bloom et al., 2007; Kobayashi et al., 2004; Popescu and Etzioni, 2005), there are two kinds of common relations between the polarity words and their targets. One is the "subject-copula-predicate" relation, such as the relationship between "long" and "battery life" in the sentence "The *battery life* of this camera is *long*". The other is the "attribute-head" relation, such as the relationship between them in the sentence "This camera has *long battery life*".

As a result, three heuristic query expansion strategies are adopted to construct efficient queries for searching. Take the collocation ⟨长,电池寿命⟩ (⟨long, battery life⟩ in English) as an example, the strategies are described as follows.

Strategy1: target + *modifier* + polarity word: Such as the query "电池寿命很长" or "电池寿命非常长" ("the battery life is *very* long" in English). Different from Strategy0, this strategy adds a *modifier* element. It refers to the words that are used to change the degree of a polarity word, such as "很" or "非常" ("very" in English). Due to the usage of the modifiers, the queries from this strategy can satisfy the "subject-copula-predicate" relation.

Strategy2: *modifier* + polarity word + 的+ target: Such as the query "很长的电池寿命" or "非常长的电池寿命" ("*very* long battery life" in English). This strategy also uses *modifiers* to modify polarity words, and the generated queries can satisfy the "attribute-head" relation.

Strategy3: *negation word* + polarity word + 的+ target: Such as the query "不长的电池寿命" or "没有长的电池寿命" ("*not* long battery life" in English). This strategy uses *negation words* to modify the polarity words. And the queries from this strategy can satisfy the "attribute-head" relation. The only difference is that the polarity of this kind of queries is opposite to that of the collocation.

Similar to the queries from Strategy0, the queries generated by Strategy1∼3 are all searched with quotes. In addition, note that the *modifier* and the *negation word* are taken from *Modifier Lexicon* and *Negation Lexicon* introduced in Table 2.

3.2.3 Pseudo Context Acquisition

For each query from Strategy0∼3, we search it in some websites to acquire the related snippets. If we directly search it using Google without site restrictions, it does return all the snippets containing the query, but lots of them are non-reviews. Further, the pseudo contexts generated by these non-reviews are useless or even harmful. To overcome this problem, the ***advanced search*** of Google is used to search the query within the *forum sites* of the product domain. We can flexibly choose several popular forum sites for each domain. The URLs of the forum sites used in this paper are listed in Table 1.

Formally, given a collocation *c*, the expanded pseudo contexts *Conx(c)* can be obtained using the following function:

$$
\begin{aligned}
Conx(c) &= \bigcup_{i=0}^{3} f(Query_i) \\
&= \bigcup_{i=0}^{3} \bigcup_{j=1}^{n} f(query_{ij})
\end{aligned}
\tag{1}
$$

Here, $Query_i$ is the query set generated by the *i*th query expansion strategy; $query_{ij}$ is the *j*th query generated by the *i*th strategy. And the parameter *n* is the total number of queries from the *i*th query expansion strategy. From this function, we can collect the contexts of *c* by summing up all the pseudo contexts from every $query_{ij}$.

Domain	URL
Camera	http://www.qqdc.com.cn
	http://forums.nphoto.net
	http://dc.pconline.com.cn
	http://photobbs.it168.com
	http://club.tech.sina.com.cn/dc
Car	http://bbs.chetx.com
	http://bbs.pcauto.com.cn
	http://club.autohome.com.cn
	http://bbs.cheshi.com
	http://www.xcar.com.cn
	http://www.autohome.com.cn
Notebook	http://benyouhui.it168.com/index.php
	http://nbbbs.zol.com.cn
	http://www.ibijiben.com
	http://notebook.pconline.com.cn
	http://nbbbs.enet.com.cn
Phone	http://bbs.imobile.com.cn
	http://sjbbs.zol.com.cn
	http://bbs.shouji.com.cn
	http://bbs.cnmo.com
	http://forum.younet.com

Table 1: The URLs used in context expansion for different domains.

In detail, the pseudo context acquisition algorithm for a collocation c is illustrated in Figure 2. Note that, the original context acquisition of c can be considered as a simplified version of the pseudo context acquisition. That's because the current review containing c can be considered as only one snippet in pseudo context acquisition. Thus, we can just carry out the two steps in (2) of Figure 2 to obtain the original contexts.

Analyzing either the pseudo contexts or the original contexts, we can find that not all of them are useful contexts. Thus we will simply filter the noisy ones by context sentiment computation, and choose the contexts showing sentiment orientations as the useful contexts.

3.3 Sentiment Analysis

For both the original and expanded pseudo contexts, we employ the lexicon-based sentiment computing method (Hu and Liu, 2004) to compute the polarity value for each context. This unsupervised approach is quite straightforward and makes use of the sentiment lexicons in Table 2.

The polarity value $Polarity(con)$ for a context con

Algorithm: Pseudo Context Expansion Algorithm

Input: A collocation c and the URL list
Output: The pseudo context set $Conx(c)$

1. Use Strategy0~3 to expand c and the expanded queries are saved as a set $Query(c)$.
2. For any query $q \in Query(c)$, acquire its pseudo contexts $Conx(q)$ as follows:
   (1) search q in the domain-related URL list, the top 100 retrieved snippets for each URL are collected as $Snip(q)$
   (2) for each snippet $sp \in Snip(q)$
   - find the sentence s containing q
   - obtain the two sentences before and after s as the contexts of q in this sp, noted as $Conx(q, sp)$
   $$Conx(q) = \bigcup_{sp \in Snip(q)} Conx(q, sp)$$
3. $$Conx(c) = \bigcup_{q \in Query(c)} Conx(q) = \bigcup_{q \in Query(c)} \bigcup_{sp \in Snip(q)} Conx(q, sp)$$

Figure 2: The algorithm for pseudo context acquisition.

Lexicon	Content
Modifier Lexicon	很, 比较, 非常, 十分, 太, 特, 特别, 挺, 相当, 格外, 分外 ("very" or "quite" in English)
Negation Lexicon	没有, 不, 不是 ("no" or "not" in English)
Positive Lexicon	There are 3,730 Chinese words are collected from HOWNET.[1]
Negative Lexicon	There are 3,116 Chinese words are collected from HOWNET.

[1] http://www.keenage.com/html/e_index.html.

Table 2: The lexicons used in this paper.

is computed by summing up the polarity values of all words in con, making use of both the word polarity defined in the positive and negative lexicons and the contextual shifters defined in the negation lexicon. The algorithm is illustrated in Figure 3.

In this algorithm, n is the parameter controlling the window size within which the negation words have influence on the polarity words, and here n is set to 3.

Normally, if the polarity value $Polarity(con)$ is more than 0, the context con is labeled as positive; if less than 0, the context is negative. We also consider the transitional words, such as "但是" ("but" in English). Finally, the contexts with positive/negative polarities are used as the useful contexts.

Domain	# of reviews	# of c (All)	# of single c (Sig)	Sig / All (%)	# of multiple c / kinds of multiple c
Camera	138	295	183	62.03	112 / 35
Car	161	232	131	56.47	101 / 33
Notebook	56	147	94	63.95	53 / 20
Phone	123	327	192	58.72	135 / 35
Total	478	1001	600	59.94	401 / 123 ≈ 3.3

Table 3: Statistics for the Chinese collocation corpus.

Algorithm: Sentiment Analysis

Input: a context *con*, and three lexicons: *Positive_Dic*, *Negative_Dic*, *Negation_Dic*

Output: Polarity value *Polarity(con)*

1. Segment *con* into word set *W(con)*
2. For each word $w \in W(con)$, compute its polarity value *Polarity(w)* as follows:
   (1) if $w \in Positive\_Dic$, *Polarity(w)* = 1;
   (2) if $w \in Negative\_Dic$, *Polarity(w)* = -1;
   (3) otherwise, *Polarity(w)* = 0;
   (4) Within the window of *n* words previous to *w*, if there is a word $w' \in Negation\_Dic$, *Polarity(w)* = -*Polarity(w)*
3. $Polarity(con) = \sum_{w \in W(con)} Polarity\ (w)$

Figure 3: The algorithm for context polarity computation.

3.4 Combination

After the pseudo context acquisition and polarity computation, two kinds of effective contexts: *original contexts* and *pseudo contexts*, and their corresponding polarities can be obtained.

In order to yield a relatively accurate polarity *Polarity(c)* for a collocation *c*, we exploit the following combination methods:

1. **Majority Voting**: Rather than considering the difference between the two kinds of contexts, this combination method relies on the polarity tag of each context. Suppose *c* has *n* effective contexts (including original and pseudo contexts), it can obtain *n* polarity tags based on the individual sentiment analysis algorithm. The polarity tag receiving more votes is chosen as the final polarity of *c*.

2. **Complementation**: For a collocation *c*, we first employ "Majority Voting" method just on the expanded pseudo contexts to obtain the polarity tag.

If the polarity of *c* cannot be recognized[2], the majority polarity tag voted on the original contexts is chosen as the final polarity tag.

4 Experimental Setup

4.1 Dataset and Evaluation Metrics

We conduct the experiments on a Chinese collocation corpus of four product domains, which is from the Task3 of the Chinese Opinion Analysis Evaluation (COAE)[3] (Zhao et al., 2008). Table 3 describes the corpus in detail.

From 478 reviews, 1,001 collocations (454 positive and 547 negative) with polarity-ambiguous words are found and manually annotated by two annotators. Cohen's kappa (Cohen, 1960), a measure of inter-annotator agreement ranging from zero to one, is 0.83, indicating a good strength of agreement [4]. In Table 3, *Sig* of the fourth column denotes the collocations that appear once in all the domain-related reviews. And *multiple* in the last column denotes the collocations that appear several times. From Table 3, we can find that among all the reviews, nearly 60% collocations only appear once. Even for the *multiple* collocations, they averagely appear less than 4 times. Therefore, for a collocation, if we only consider its original contexts alone or the expanded pseudo contexts from the domain-related review set alone, the contexts are obviously limited and unreliable.

Instead of using accuracy, we use precision (*P*), recall (*R*) and F-measure (*F1*) to measure the performance of this task. That's because two kinds of collocations' polarities cannot be disambiguated. One

---

[2] The reason will be explained in the last paragraph of Section 4.1.

[3] http://www.ir-china.org.cn/coae2008.html

[4] A small number of collocations are still difficult to be disambiguated from contexts.

is the sparse collocations, which obtain no effective contexts. The other is the collocations that acquire the same amount of positive and negative contexts. The metrics are defined as follows.

$$P = \frac{correctly\ disambiguated\ collocations}{disambiguated\ collocations} \quad (2)$$

$$R = \frac{correctly\ disambiguated\ collocations}{all\ collocations} \quad (3)$$

$$F1 = \frac{2PR}{P+R} \quad (4)$$

4.2 System Description

In order to compare our method with previous work, we build several systems as follows:

$NoExp$: Following the method proposed by Hu (Hu and Liu, 2004), without using the expanded pseudo contexts, we only consider the two original contexts Sen_{bef} and Sen_{aft} of a collocation c in the current review. **If** Sen_{bef} expresses the polarity *polar*, then $Polarity(ac) = polar$. **Else if** Sen_{aft} expresses the polarity $polar'$, then $Polarity(ac) = polar'$. **Else**, this method cannot disambiguate the polarity of c. In this method, the transitional words, such as "但是" ("but" in English) are considered.

$Exp_{dataset}$: Following the method proposed by Ding (Ding et al., 2008), we solve this task with the help of the pseudo contexts in the domain-related review dataset. For a collocation c appearing in many domain-related reviews, this method refers to the polarities of the same c in other reviews. The majority polarity is chosen as final polarity.

$Exp_{web+sig}$: This method is the same as our method in this paper, except for (1) not combining the original contexts, and (2) not using all the three query expansion strategies, but just using the *single* (abbv. *sig*) Strategy0. This method expands the pseudo contexts from the *web*. The majority polarity is chosen as the final polarity.

$Exp_{web+exp}$: This method is the same as our proposed method in this paper, except for not combining the original contexts. It expands the pseudo contexts from the *web*. And the "*exp*" in the subscript means that this method uses all the query expansion strategies. The majority polarity of all the pseudo contexts is chosen as the final polarity.

$Exp_{web+exp+com}^{mv/c}$: This is the method proposed in this paper, which combines the original and expanded pseudo contexts. The superscript "*mv/c*" is short for the two combination methods: **M**ajority **V**oting and **C**omplementation.

5 Results

5.1 Comparisons among All the Systems

In fact, all the systems shown in Section 4.2 can be considered as *context* based methods. The essential difference among them lies in the contexts they used. For a collocation, the contexts for $NoExp$ are two original contexts from the current review. Breaking down the boundary of the current review, $Exp_{dataset}$ explores the pseudo contexts from other domain-related reviews. Further, $Exp_{web+sig}$, $Exp_{web+exp}$ and $Exp_{web+exp+com}^{mv/c}$ expand the pseudo contexts from web, which can be considered as a large corpus and can provide more evidences for the collocation polarity disambiguation.

System	$P(\%)$	$R(\%)$	$F1(\%)$
$NoExp$	67.32	41.16	51.08
$Exp_{dataset}$	68.14	47.85	56.22
$Exp_{web+sig}$	70.00	53.85	60.87
$Exp_{web+exp}$	74.97	63.14	68.55
$Exp_{web+exp+com}^{mv}$	75.53	67.83	71.47
$Exp_{web+exp+com}^{c}$	74.36	69.83	72.02

Table 4: Comparative results for the collocation polarity disambiguation task.

Table 4 illustrates the comparative results of all systems for collocation polarity disambiguation. It can be observed that our system $Exp_{web+exp+com}^{mv}$ and $Exp_{web+exp+com}^{c}$ outperform all the other systems. We discuss the experimental results as follows:

$NoExp$ yields the worst performance, especially on the recall. The reason is that the original contexts used in this system are limited, and some of them are even noisy. In comparison, $Exp_{dataset}$ adds a post-processing step of expanding pseudo contexts from the topically-related review dataset, which achieves a better result with an absolute improvement of 5.14% (*F1*). This suggests that the contexts expanded from other reviews are helpful in disambiguating the collocation's polarity.

166

However, $Exp_{dataset}$ is just effective in disambiguating the polarity of such a collocation c, which appears many times in the domain-related reviews. From Table 3, we can notice that this kind of collocations only accounts for 40% in all the collocations, and further they appear less than 4 times on average. Thus, for such a collocation c, the pseudo contexts expanded from other reviews that contain the same c are still far from enough, since the review set size in this system is not very large.

In order to avoid the context limitation problem, we expand more pseudo contexts from web for each collocation. We first try to use a simple query form (Strategy0) for web mining. Table 4 illustrates that the corresponding system $Exp_{web+sig}$ outperforms the system $Exp_{dataset}$. It can demonstrate that our web mining based pseudo context expansion is useful for disambiguating the collocation's polarity, since this system can explore more contexts. However, we can find that the performance is not very ideal. This system can generate some harmful contexts for the reason of the wrong modifying relations between polarity words and targets in the retrieved snippets.

Thus this paper adds three query expansion strategies to generate more and accurate pseudo contexts. Table 4 shows that the corresponding system $Exp_{web+exp}$ can achieve a better result with $F1$ = 68.55%, which is significantly (χ^2 test with $p <$ 0.01) outperforms $Exp_{web+sig}$. It demonstrates that the query expansion strategies are useful.

Finally, Table 4 gives the results of our method in this paper, $Exp_{web+exp+com}^{mv}$ and $Exp_{web+exp+com}^{c}$, which combines the original and expanded pseudo contexts to yield a final polarity. We can observe that both of these systems outperform the system $NoExp$ of just using the original contexts and the system $Exp_{web+exp}$ of just using the expanded pseudo contexts. This can illustrate that the two kinds of contexts are complementary to each other. In addition, we can also find that the two combination methods produce similar results. In detail, $Exp_{web+exp+com}^{mv}$ disambiguates 899 collocations, 679 of them are correct; $Exp_{web+exp+com}^{c}$ disambiguates 940 collocations, 699 of them are correct.

We can further find that, although the amount of original contexts is small, it also plays an important role in disambiguating the polarities of the collocations that cannot be recognized by the expanded pseudo contexts.

5.2 The Contributions of the Query Expansion Strategies

The expanded pseudo contexts from our method can be partly credited to the query expansion strategies. Based on this, this section aims to analyze the different contributions of the query expansion strategies in our method.

Strategy	P(%)	R(%)	F1(%)	Avg(#)
Strategy0	70.00	53.85	60.87	71
Strategy1	74.14	55.84	63.70	112
Strategy2	61.84	37.56	46.74	26
Strategy3	64.34	33.17	43.77	20
$Exp_{web+exp}$	74.97	63.14	**68.55**	229

Table 5: The performance of our method based on each query expansion strategy for collocation polarity disambiguation.

Table 5 provides the performance of our method based on each query expansion strategy for collocation polarity disambiguation. For each strategy, "*Avg*" in Table 5 denotes the average number of the expanded pseudo contexts for each collocation. From this table, we can find that the larger the "*Avg*" is, the better (*F1*) the strategy is. In detail, Strategy1 with the largest "*Avg*" has the best performance; and Strategy3 with the fewest "*Avg*" has the worst performance. This can further demonstrate our idea that more and effective pseudo contexts can improve the performance of the collocation polarity disambiguation task. $Exp_{web+exp}$ integrates all the query expansion strategies and obtains much more "*Avg*". Therefore, this can significantly increase the recall value, and further produce a better result. On the other hand, the results in Table 5 show that these heuristic query expansion strategies are effective.

5.3 Deep Experiments in the Three-Component Framework

In order to do a detailed analysis into our three-component framework, some deep experiments are made:

Query Expansion The aim of query expansion is to retrieve lots of relative snippets, from which we can extract the useful pseudo contexts. For each

	Strategy0 (%)	Strategy1 (%)	Strategy2 (%)	Strategy3 (%)
Query Expansion	76.75	94.50	85.50	85.25
Pseudo Context	71.25	73.50	67.50	74.50
Sentiment Analysis	63.00	68.25	59.00	69.75

Table 6: The accuracies of the query expansion, pseudo context and sentiment analysis for each strategy.

snippet, if the polarity word of the collocation does modify the target, we consider this snippet as a correct query expansion result.

Pseudo Context For each expanded pseudo context from web, if it shows the same sentiment orientation with the collocation (or opposite with the collocation's polarity because of the usage of transitional words), we consider this context as a correct pseudo context.

Sentiment Analysis For each expanded pseudo context, if its polarity can be correctly recognized by the polarity computation method in Figure 3, and meanwhile it shows the same sentiment orientation with the collocation, we consider this context as a correct one.

Table 6 illustrates the accuracy of each experiment for each strategy in detail, where 400 web retrieved snippets for *Query Expansion* and 400 expanded pseudo contexts for *Pseudo Context* and *Sentiment Analysis* are randomly selected and manually evaluated for each strategy.

Seen from Table 6, we can find that:

1. For *Query Expansion*, all strategies yield good accuracies except for Strategy0. This can draw a same conclusion with our analysis in Section 3.2.1. The queries from Strategy0 are short, thus in many retrieved snippets, there exist no modifying relations between the polarity words and targets. Accordingly, the pseudo contexts from these snippets are incorrect. This can result in the low accuracy of Strategy0. On the other hand, we can find that the other three query expansion strategies perform well.

2. Although the final result of our three-component framework is good, the accuracies of *Pseudo Context* and *Sentiment Analysis* for each strategy is not very high. This is perhaps caused by unrefined work on the specific sub-stages. For example, we get all the pseudo contexts using the algorithm in Figure 2. However, in some reviews, the two sentences before and after the target sentence

have no polarity relation with the target sentence itself. This can bring in some noises. On the other hand, the context polarity computation algorithm in Figure 3 is just a simple attempt, which is not the best way to compute the context's polarity.

In fact, this paper aims to try some simple algorithms for each component to validate the effectiveness of the three-component framework. We will polish every component of our framework in future.

6 Conclusion and Future Work

This paper proposes a web-based context expansion framework for collocation polarity disambiguation. The basic assumption of this framework is that, if a collocation appears in different forms, both within the same review and within topically-related reviews, then the large amounts of pseudo contexts from these reviews can help to disambiguate such a collocation's polarity. Based on this assumption, this framework includes three independent components. First, the heuristic query expansion strategies are adopted to expand pseudo contexts from web; then a simple but effective polarity computation method is used to recognize the polarities for both the original contexts and the expanded pseudo contexts; and finally, we integrate the polarities from the original and pseudo contexts as the collocation's polarity. Without using any additional labeled data, experiments on a Chinese data set from four product domains show that the proposed framework outperforms other previous work.

This paper can be concluded as follows:

1. A framework including three independent components is proposed for collocation polarity disambiguation. We can try other different algorithms for each component.

2. Web-based pseudo contexts are effective for disambiguating a collocation's polarity.

3. The query expansion strategies are promising, which can generate more useful and correct contexts.

4. The initial contexts from current reviews and the expanded contexts from web are complementary to each other.

The immediate extension of our work is to polish each component of this framework, such as improving the accuracy of query expansion and pseudo context acquisition, using other effective polarity computing methods for each context and so on. In addition, we will explore other query expansion strategies to generate more effective contexts.

Acknowledgments

References

Kenneth Bloom, Navendu Garg, and Shlomo Argamon. 2007. Extracting appraisal expressions. In *HLT-NAACL 2007*, pages 308–315.

D. Bollegala, D. Weir, and J. Carroll. 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 132–141. Association for Computational Linguistics.

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

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the Conference on Web Search and Web Data Mining (WSDM)*, pages 231–240.

A. Esuli and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss analysis. In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, pages 617–624.

A. Esuli. 2008. Automatic generation of lexical resources for opinion mining: models, algorithms and applications. In *ACM SIGIR Forum*, volume 42, pages 105–106. ACM.

V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics.

Yulan He, Chenghua Lin, and Harith Alani. 2011. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 123–131, Portland, Oregon, USA, June. Association for Computational Linguistics.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

V. Jijkoun, M. De Rijke, and W. Weerkamp. 2010. Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594. Association for Computational Linguistics.

N. Kaji and M. Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083.

Jaap Kamps, Maarten Marx, R. ort. Mokken, and Maarten de Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of LREC-2004*, pages 1115–1118.

H. Kanayama and T. Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics.

Soo-Min Kim and Eduard Hovy. 2005. Automatic detection of opinion bearing words and sentences. In *Proceedings of IJCNLP-2005*, pages 61–66.

S.-M. Kim and E. Hovy. 2006. Identifying and analyzing judgment opinions. In *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, pages 200–207.

Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 584–589.

Binyang Li, Lanjun Zhou, Shi Feng, and Kam-Fai Wong. 2010. A unified graph model for sentence-based opinion retrieval. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, page 1367 – 1375.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of WWW-2005*, pages 342–351.

Y. Lu, M. Castellanos, U. Dayal, and C.X. Zhai. 2011. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM.

S. Mohammad, C. Dunne, and B. Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 599–608. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP-2002*, pages 79–86.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *hltemnlp2005*, pages 339–346.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP-2003*, pages 105–112.

Ellen Riloff, Janyce Wiebe, and William Phillips. 2005. Exploiting subjectivity classification to improve information extraction. In *Proceedings of AAAI-2005*, pages 1106–1111.

Fangzhong Su and Katja Markert. 2009. Subjectivity recognition on word senses via semi-supervised mincuts. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 1–9.

Yasuhiro Suzuki, Hiroya Takamura, and Manabu Okumura. 2006. Application of semi-supervised learning to evaluative expression classification. In *Computational Linguistics and Intelligent Text Processing*, pages 502–513.

P. Turney, M.L. Littman, et al. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785.

Jan Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL)*, pages 1065–1072.

Janyce Wiebe, Eric Breck, and Chris Buckley. 2003. Recognizing and Organizing Opinions Expressed in the World Press. In *Papers from the AAAI Spring Symposium on New Directions in Question Answering*, pages 24–26.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of AAAI*, pages 735–740.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP-2005*, pages 347–354.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3).

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP-2003*, pages 129–136.

Min Zhang and Xingyao Ye. 2008. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, pages 411–419.

Wei Zhang, Clement Yu, and Weiyi Meng. 2007. Opinion retrieval from blogs. In *In proceedings of CIKM*, page 831 – 840.

Jun Zhao, Hongbo Xu, Xuanjing Huang, Songbo Tan, Kang Liu, and Qi Zhang. 2008. Overview of chinese opinion analysis evaluation 2008.