Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities

Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, Kam-Fai Wong
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, NT, Hong Kong, China
Key Laboratory of High Confidence Software Technologies
Ministry of Education, China
{ljzhou, byli, wgao, zywei, kfwong}@se.cuhk.edu.hk

Abstract

Polarity classification of opinionated sentences with both positive and negative sentiments[1] is a key challenge in sentiment analysis. This paper presents a novel unsupervised method for discovering intra-sentence level discourse relations for eliminating polarity ambiguities. Firstly, a discourse scheme with discourse constraints on polarity was defined empirically based on Rhetorical Structure Theory (RST). Then, a small set of cue-phrase-based patterns were utilized to collect a large number of discourse instances which were later converted to semantic sequential representations (SSRs). Finally, an unsupervised method was adopted to generate, weigh and filter new SSRs without cue phrases for recognizing discourse relations. Experimental results showed that the proposed methods not only effectively recognized the defined discourse relations but also achieved significant improvement by integrating discourse information in sentence-level polarity classification.

1 Introduction

As an important task of sentiment analysis, polarity classification is critically affected by discourse structure (Polanyi and Zaenen, 2006). Previous research developed discourse schema (Asher et al., 2008) (Somasundaran et al., 2008) and proved that the utilization of discourse relations could improve the performance of polarity classification on dialogues (Somasundaran et al., 2009). However, current state-of-the-art methods for sentence-level polarity classification are facing difficulties in ascertaining the polarity of some sentences. For example:

(a) [Although Fujimori was criticized by the international community]，[he was loved by the domestic population]，[because people hated the corrupted ruling class].（儘管國際間對藤森口誅筆伐，他在國內一直深受百姓愛戴，原因是百姓對腐化的統治階級早就深惡痛絕。）

Example (a) is a positive sentence holding a *Contrast* relation between first two segments and a *Cause* relation between last two segments. The polarity of "criticized", "hated" and "corrupted" are recognized as negative expressions while "loved" is recognized as a positive expression. Example (a) is difficult for existing polarity classification methods for two reasons: (1) the number of positive expressions is less than negative expressions; (2) the importance of each sentiment expression is unknown. However, consider Figure 1, if we know that the polarity of the first two segments holding a *Contrast* relation is determined by the *nucleus* (Mann and Thompson, 1988) segment and the polarity of the last two segments holding a *Cause* relation is also determined by the *nucleus* segment, the polarity of the sentence will be determined by the polarity of "[he...population]". Thus, the polarity of Example (a) is positive.

Statistics showed that 43% of the opinionated sentences in NTCIR[2] MOAT (Multilingual Opinion Analysis Task) Chinese corpus[3] are *ambiguous*. Existing sentence-level polarity classification methods ignoring discourse structure often give wrong results for these sentences. We implemented state-of-the-

[1] Defined as *ambiguous* sentences in this paper

[2] http://research.nii.ac.jp/ntcir/

[3] Including simplified Chinese and traditional Chinese corpus from NTCIR-6 MOAT and NTCIR-7 MOAT

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 162–171,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

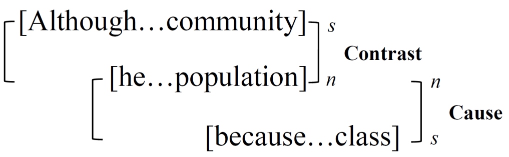

Figure 1: Discourse relations for Example (a). (*n* and *s* denote *nucleus* and *satellite* segment, respectively)

art method (Xu and Kit, 2010) in NTCIR-8 Chinese MOAT as the baseline polarity classifier (BPC) in this paper. Error analysis of BPC showed that 49% errors came from *ambiguous* sentences.

In this paper, we focused on the automation of recognizing intra-sentence level discourse relations for polarity classification. Based on the previous work of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), a discourse scheme with discourse constraints on polarity was defined empirically (see Section 3). The scheme contains 5 relations: *Contrast*, *Condition*, *Continuation*, *Cause* and *Purpose*. From a raw corpus, a small set of cue-phrase-based patterns were used to collect discourse instances. These instances were then converted to semantic sequential representations (SSRs). Finally, an unsupervised SSR learner was adopted to generate, weigh and filter high quality new SSRs without cue phrases. Experimental results showed that the proposed methods could effectively recognize the defined discourse relations and achieve significant improvement in sentence-level polarity classification comparing to BPC.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 presents the discourse scheme with discourse constraints on polarity. Section 4 gives the detail of proposed method. Experimental results are reported and discussed in Section 5 and Section 6 concludes this paper.

2 Related Work

Research on polarity classification were generally conducted on 4 levels: document-level (Pang et al., 2002), sentence-level (Riloff et al., 2003), phrase-level (Wilson et al., 2009) and feature-level (Hu and Liu, 2004; Xia et al., 2007).

There was little research focusing on the automatic recognition of intra-sentence level discourse relations for sentiment analysis in the literature. Polanyi and Zaenen (2006) argued that valence calculation is critically affected by discourse structure. Asher et al. (2008) proposed a shallow semantic representation using a feature structure and use five types of rhetorical relations to build a fine-grained corpus for deep contextual sentiment analysis. Nevertheless, they did not propose a computational model for their discourse scheme. Snyder and Barzilay (2007) combined an agreement model based on contrastive RST relations with a local aspect model to make a more informed overall decision for sentiment classification. Nonetheless, contrastive relations were only one type of discourse relations which may help polarity classification. Sadamitsu et al. (2008) modeled polarity reversal using HCRFs integrated with inter-sentence discourse structures. However, our work is on intra-sentence level and our purpose is not to find polarity reversals but trying to adapt general discourse schemes (e.g., RST) to help determine the overall polarity of *ambiguous* sentences.

The most closely related works were (Somasundaran et al., 2008) and (Somasundaran et al., 2009), which proposed *opinion frames* as a representation of discourse-level associations on dialogue and modeled the scheme to improve opinion polarity classification. However, *opinion frames* was difficult to be implemented because the recognition of opinion target was very challenging in general text. Our work differs from their approaches in two key aspects: (1) we distinguished *nucleus* and *satellite* in discourse but *opinion frames* did not; (2) our method for discourse discovery was unsupervised while their method needed annotated data.

Most research works about discourse classification were not related to sentiment analysis. Supervised discourse classification methods (Soricut and Marcu, 2003; Duverle and Prendinger, 2009) needed manually annotated data. Marcu and Echihabi (2002) presented an unsupervised method to recognize discourse relations held between arbitrary spans of text. They showed that lexical pairs extracted from massive amount of data can have a major impact on discourse classification. Blair-Goldensohn et al. (2007) extended Marcu's work by using parameter opitimization, topic segmentation and syntactic parsing. However, syntactic parsers were usually costly and impractical when dealing with large scale of text. Thus, in additional to lexical features, we incorporated sequential and semantic information in proposed method for discourse relation classification. Moreover, our method kept the characteristic of language independent, so it could be applied to other languages.

3 Discourse Scheme for Eliminating Polarity Ambiguities

Since not all of the discourse relations in RST would help eliminate polarity ambiguities, the discourse scheme defined in this paper was on a much coarser level. In order to ascertain which relations should be included in our scheme, 500 *ambiguous* sentences were randomly chosen from NTCIR MOAT Chinese corpus and the most common discourse relations for connecting independent clauses in compound sentences were annotated. We found that 13 relations from RST occupied about 70% of the annotated discourse relations which may help eliminate polarity ambiguities. Inspired by Marcu and Echihabi (2002), to construct relatively low-noise discourse instances for unsupervised methods using cue phrases, we grouped the 13 relations into the following 5 relations:

Contrast is a union of *Antithesis*, *Concession*, *Otherwise* and *Contrast* from RST.

Condition is selected from RST.

Continuation is a union of *Continuation*, *Parallel* from RST.

Cause is a union of *Evidence*, *Volitional-Cause*, *Nonvolitional-Cause*, *Volitional-result* and *Nonvolitional-result* from RST.

Purpose is selected from RST.

The discourse constraints on polarity presented here were based on the observation of annotated discourse instances: (1) discourse instances holding *Contrast* relation should contain two segments with opposite polarities; (2) discourse instances holding *Continuation* relation should contain two segments with the same polarity; (3) the polarity of discourse instances holding *Contrast*, *Condition*, *Cause* or *Purpose* was determined by the *nucleus* segment; (4) the polarity of discourse instances holding *Continuation* was determined by either segment.

Relation	Cue Phrases (English Translation)
Contrast	although^{1}, but^{2}, however^{2}
Condition	if^{1}, (if^{1}， then^{2})
Continuation	and, further more, (not only, but also)
Cause	because^{1}, thus^{2}, accordingly^{2}, as a result^{2}
Purpose	in order to^{2}, in order that^{2}, so that^{2}

^{1} means *CUE1* and ^{2} means *CUE2*

Table 1: Examples of cue phrases

4 Methods

The proposed methods were based on two assumptions: (1) Cue-phrase-based patterns could be used to find limited number of high quality discourse instances; (2) discourse relations were determined by lexical, structural and semantic information between two segments.

Cue-phrase-based patterns could find only limited number of discourse instances with high precision (Marcu and Echihabi, 2002). Therefore, we could not rely on cue-phrase-based patterns alone. Moreover, there was no annotated corpus similar to Penn Discourse TreeBank (Miltsakaki et al., 2004) in other languages such as Chinese. Thus, we proposed a language independent unsupervised method to identify discourse relations without cue phrases while maintaining relatively high precision. For each discourse relation, we started with several cue-phrase-based patterns and collected a large number of discourse instances from raw corpus. Then, discourse instances were converted to semantic sequential representations (SSRs). Finally, an unsupervised method was adopted to generate, weigh and filter common SSRs without cue phrases. The mined common SSRs could be directly used in our SSR-based classifier in unsupervised manner or be employed as effective features for supervised methods.

4.1 Gathering and representing discourse instances

A discourse instance, denoted by D_i, consists of two successive segments ($D_{i[1]}, D_{i[2]}$) within a sentence. For example:

D_1: [*Although Boris is very brilliant at math*]$_s$, [*he*

BOS... ， *[CUE2]*...EOS
BOS *[CUE1]*... ， ...EOS
BOS... ， *[CUE1]*...EOS
BOS *[CUE1]*... ， *[CUE2]*...EOS

Table 2: Cue-phrase-based patterns. BOS and EOS denoted the beginning and end of two segments.

is a horrible teacher]$_n$
D_2: [*John is good at basketball*]$_s$, [*but he lacks team spirit*]$_n$

In D_1, "although" indicated the *satellite* section while in D_2, "but" indicated the *nucleus* section. Accordingly, different cue phrases may indicate different segment type. Table 1 listed some examples of cue phrases for each discourse relation. Some cue phrases were singleton (e.g. "although" and "as a result") and some were used as a pair (e.g. "not only, but also"). "*CUE1*" indicated *satellite* segments and "*CUE2*" indicated *nucleus* segments. Note that we did not distinguish *satellite* from *nucleus* for *Continuation* in this paper because the polarity could be determined by either segment.

Table 2 listed cue-phrase-based patterns for all relations. To simplify the problem of discourse segmentation, we split compound sentences into discourse segments using commas and semicolons. Although we collected discourse instances from compound sentences only, the number of instances for each discourse relation was large enough for the proposed unsupervised method. Note that we only collected instances containing at least one sentiment word in each segment.

In order to incorporate lexical and semantic information in our method, we represented each word in a discourse instance using a part-of-speech tag, a semantic label and a sentiment tag. Then, all discourse instances were converted to SSRs. The rules for converting were as follows:

(1) Cue phrases and punctuations were ingored. But the information of *nucleus*(n) and *satellite*(s) was preserved.

(2) Adverbs(*RB*) appearing in sentiment lexicon, verbs(*V*), adjectives(*JJ*) and nouns(*NN*) were represented by their part-of-speech (*pos*) tag with semantic label ($semlabel$) if available.

(3) Named entities (*NE*; *PER*: person name; *ORG*: organization), pronouns (*PRP*), and function words were represented by their corresponding named entity tags and part-of-speech tags, respectively.

(4) Added sentiment tag (P: Positive; N: Negative) to all sentiment words.

By applying above rules, the SSRs for D_1 and D_2 would be:

d_1: [*PER V|Ja01 RB|Ka01 JJ|Ee14|P IN NN|Dk03*]$_s$, [*PRP V|Ja01 DT JJ|Ga16|N NN|Ae13*]$_n$
d_2: [*PER V|Ja01 JJ|Ee14|P IN NN|Bp12*]$_s$, [*PRP V|He15|N NN|Di10 NN|Dd08*]$_n$

Refer to d_1 and d_2, "Boris" could match "John" in SSRs because they were converted to "*PER*" and they all appeared at the beginning of discourse instances. "*Ja01*", "*Ee14*" etc. were semantic labels from Chinese synonym list extended version (Che et al., 2010). There were similar resources in other languages such as Wordnet(Fellbaum, 1998) in English. The next problem became how to start from current SSRs and generate new SSRs for recognizing discourse relations without cue phrases.

4.2 Mining common SSRs

Recall assumption (2), in order to incorporate lexical, structural and semantic information for the similarity calculation of two SSRs holding the same discourse relation, three types of matches were defined for $\{(u, v)|u \in d_{i[k]}, v \in d_{j[k]}, k = 1, 2\}$: (1)Full match: (i) $u = v$ or (ii) $u.pos = v.pos$ and $u.semlabel = v.semlabel$ or (iii) $u.pos = v.pos$ and u had a sentiment tag and v had a sentiment tag or (iv) $u.pos$ and $v.pos \in \{$*PRP, PER, ORG*$\}$ (2) Partial match: $u.pos = v.pos$ but not Full match; (3) Mismatch: $u.pos \neq v.pos$.

Generating common SSRs

Intuitively, a simple way of estimating the similarity between two SSRs was using the number of mismatches. Therefore, we utilized $match(d_i, d_j)$ where $i \neq j$, which integrated the three types of matches defined above to calculate the number of mismatches and generate common SSRs. Consider Table 3, in common SSRs, full matches were preserved, partial matches were replaced by part of speech tags and mismatches were replaced by '*'s. The common SSRs generated during the calculation of $match(d_i, d_j)$ consisted of two parts. The first part was generated by $d_{i[1]}$ and $d_{j[1]}$ and the second part was generated by $d_{i[2]}$ and $d_{j[2]}$. We stipulated

d_1	d_2	mis	$conf$	ssr
PER	*PER*	0	0	*PER*
V|Ja01	*V|Ja01*	0	0	*V|Ja01*
RB|Ka01		+1	-0.298	*
JJ|Ee14|P	*JJ|Ee14|P*	0	0	*JJ|Ee14|P*
IN	*IN*	0	0	*IN*
NN|Dk03	*NN|Bp12*	0	-0.50	*NN*
$conf(ssr_{[1]}) = -0.798$				
PRP	*PRP*	0	0	*PRP*
V|Ja01	*V|He15|N*	0	-0.50	*V*
DT		+1	-0.184	*
JJ|Ga16|N		+1	-1.0	*
NN|Ae13	*NN|Di10*	0	-0.50	*NN*
	NN|Dd08	+1	-1.0	*
$conf(ssr_{[2]}) = -3.184$				

Table 3: Calculation of $match(d_1, d_2)$. ssr denoted the common SSR between d_1 and d_2 , $conf(ssr_{[1]})$ and $conf(ssr_{[2]})$ denoted the confidence of ssr.

that d_i and d_j could generate a common SSR if and only if the orders of *nucleus* segment and *satellite* segment were the same.

In order to guarantee relatively high quality common SSRs, we empirically set the upper threshold of the number of mismatches as 0.5 (i.e., $\leq 1/2$ of the number of words in the generated SSR). It's not difficult to figure out that the number of mismatches generated in Table 3 satisfied this requirement. As a result, for each discourse relation r_n, a corresponding common SSR set S_n could be obtained by adopting $match(d_i, d_j)$ where $i \neq j$ for all discourse instances. An advantage of $match(d_1, d_2)$ was that the generated common SSRs preserved the sequential structure of original discourse instances. And common SSRs allows us to build high precision discourse classifiers (See Section 5).

Weighing and filtering common SSRs

A problem of $match(d_i, d_j)$ was that it ignored some important information by treating different mismatches equally. For example, the adverb "very" in "very brilliant" of D_1 was not important for discourse recognition. In other words, the number of mismatches in $match(d_i, d_j)$ could not precisely reflect the confidence of the generated common SSRs. Therefore, it was needed to weigh different mismatches for the confidence calculation of common SSRs.

Intuitively, if a partial match or a mismatch (denoted by u_m) occurred very frequently in the generation of common SSRs, the importance of u_m tends to diminish. Inspired by the *tf-idf* model, given $ssr_i \in S_n$, we utilized the following equation to estimate the weight (denoted by w_m) of u_m.

$$w_m = -uf_m \cdot \log\left(|S_n|/ssrf_m\right)$$

where uf_m denoted the frequency of u_m during the generation of ssr_i, $|S_n|$ denoted the size of S_n and $ssrf_m$ denoted the number of common SSRs in S_n containing u_m . All weights were normalized to $[-1, 0)$.

Nouns (except for named entities) and verbs were most representative words in discourse recognition (Marcu and Echihabi, 2002). In addition, adjectives and adverbs appearing in sentiment lexicons were important for polarity classification. Therefore, for these 4 kinds of words, we utilized -1.0 for a mismatch and -0.50 for a partial match.

As we had got the weights for all partial matches and mismatches, the confidence of $ssr_i \in S_n$ could be calculated using the cumulation of weights of partial matches and mismatches in $ssr_{i[1]}$ and $ssr_{i[2]}$. Recall Table 3, $conf(ssr_{[1]})$ and $conf(ssr_{[2]})$ represented the confidence scores of $match(d_{i[1]}, d_{j[1]})$ and $match(d_{i[2]}, d_{j[2]})$, respectively. In order to control the quantity and quality of mined SSRs, a threshold $minconf$ was introduced. ssr_i will be preserved if and only if $conf(ssr_{i[1]}) \geq minconf$ and $conf(ssr_{i[2]}) \geq minconf$. The value of $minconf$ was tuned using the development data.

Finally, we combined adjacent '*'s and preserved SSRs containing at least one notional word and at least two words in each segment to meet the demand of maintaining high precision (e.g., "[* *DT* *]", "[*PER* *]" will be dropped). Moreover, since many of the SSRs were duplicated, we ranked all the generated SSRs according to their occurrences and dropped those appearing only once in order to preserve common SSRs. At last, SSRs appearing in more than one common SSR set were removed for maintaining the uniqueness of each set. The common SSR set S_n for each discourse relation r_n could be directly used in SSR-based unsupervised classifiers or be employed as effective features in supervised methods.

Relation	Occurrence
Contrast	86 (8.2%)
Condition	27 (2.6%)
Continuation	445 (42.2%)
Cause	123 (11.7%)
Purpose	55 (5.2%)
Others	318 (30.2%)

Table 4: Distribution of discourse relations on NTC-7. *Others* represents discourse relations not included in our discourse scheme.

5 Experiments

5.1 Annotation work and Data

We extracted all compound sentences which may contain the defined discourse relations from opinionated sentences (neutral ones were dropped) of NTCIR7 MOAT simplified Chinese training data. 1,225 discourse instances were extracted and two annotators were trained to annotate discourse relations according to the discourse scheme defined in Section 3. Note that we annotate both explicit and implicit discourse relations. The overall inter annotator agreement was 86.05% and the *Kappa-value* was 0.8031. Table 4 showed the distribution of annotated discourse relations based on the inter-annotator agreement. The proportion of occurrences of each discourse relations varied greatly. For example, *Continuation* was the most common relation in annotated corpus, but the occurrences of *Condition* relation were rare.

The experiments of this paper were performed using the following data sets:

NTC-7 contained manually annotated discourse instances (shown in Table 4). The experiments of discourse identification were performed on this data set.

NTC-8 contained all opinionated sentences (neutral ones were dropped) extracted from NTCIR8 MOAT simplified Chinese test data. The experiments of polarity ambiguity elimination using the identified discourse relations were performed on this data set.

XINHUA contained simplified Chinese raw news text from Xinhua.com (2002-2005). A word segmentation tool, a part-of-speech tagging tool, a named entity recognizer and a word sense disambiguation tool (Che et al., 2010) were adopted to all sentences. The common SSRs were mined from this data set.

5.2 Experimental Settings

Discourse relation identification

In order to systematically justify the effectiveness of proposed unsupervised method, following experiments were performed on NTC-7:

Baseline used only cue-phrase-based patterns.

M&E proposed by Marcu and Echihabi (2002). Given a discourse instance D_i, the probabilities: $P(r_k|(D_{i[1]}, D_{i[2]}))$ for each relation r_k were estimated on all text from XINHUA. Then, the most likely discourse relation was determined by taking the maximum over $argmax_k\{P(r_k|(D_{i[1]}, D_{i[2]})\}$.

cSSR used both cue-phrase-based patterns together with common SSRs for recognizing discourse relations. Common SSRs were mined from discourse instances extracted from XINHUA using cue-phrase-based patterns. Development data were randomly selected for tuning $minconf$.

SVM was trained utilizing cue phrases, probabilities from *M&E*, topic similarity, structure overlap, polarity of segments and mined common SSRs (Optional). The parameters of the SVM classifier were set by a grid search on the training set. We performed 4-fold cross validation on NTC-7 to get an average performance.

The purposes of introducing *SVM* in our experiment were: (1) to compare the performance of *cSSR* to supervised method; (2) to examine the effectiveness of integrating common SSRs as features for supervised methods.

Polarity ambiguity elimination

BPC was trained mainly utilizing punctuation, uni-gram, bi-gram features with confidence score output. Discourse classifiers such as *Baseline*, *cSSR* or *SVM* were adopted individually for the post-processing of BPC. Given an *ambiguous* sentence which contained more than one segment, an intuitive three-step method was adopted to integrated a discourse classifier and discourse constraints on polarity for the post-processing of BPC:

(1) Recognize all discourse relations together with *nucleus* and *satellite* information using a discourse classifier. The *nucleus* and *satellite* information is

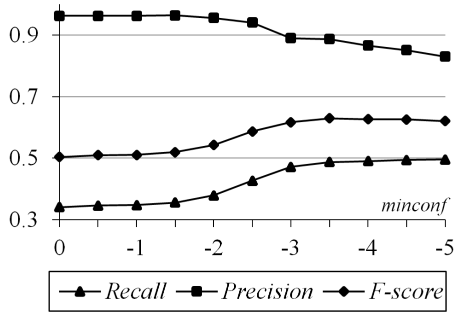

Figure 2: Influences of different values of $minconf$ to the performance of cSSR

acquired by cSSR if a segment pair could match a cSSR. Otherwise, we use the annotated *nucleus* and *satellite* information.

(2) Apply discourse constraints on polarity to ascertain the polarity for each discourse instance. There may be conflicts between polarities acquired by BPC and discourse constraints on polarity (e.g., Two segments with the same polarity holding a *Contrast* relation). To handle this problem, we chose the segment with higher polarity confidence and adjusted the polarity of the other segment using discourse constraints on polarity.

(3) If there was more than one discourse instance in a single sentence, the overall polarity of the sentence was determined by voting of polarities from each discourse instance under the majority rule.

5.3 Experimental Results

Refer to Figure 2, the performance of *cSSR* was significantly affected by $minconf$. Note that we performed the tuning process of $minconf$ on different development data ($1/4$ instances randomly selected from NTC-7) and Figure 2 showed the average performance. *cSSR* became *Baseline* when $minconf = 0$. A significant drop of precision was observed when $minconf$ was less than -2.5. The recall remained around 0.495 when $minconf \leq -4.0$. The best performance was observed when $minconf = -3.5$. As a result, -3.5 was utilized as the threshold value for *cSSR* in the following experiments.

Table 5 presented the experimental results for discourse relation classification. it showed that:

(1) Cue-phrase-based patterns could find only limited number of discourse relations (34.1% of average recall) with a very high precision (96.17% of average precision). This is a proof of assumption (1) given in Section 4. On the other side, *M&E* which only considered word pairs between two segments of discourse instances got a higher recall with a large drop of precision. The drop of precision may be caused by the neglect of structural and semantic information of discourse instances. However, *M&E* still outperformed *Baseline* in average F-*score*.

	BPC	*Baseline*	*cSSR*	*SVM* +SSRs
Precision	0.7661	0.7982	0.8059	**0.8113**
Recall	0.7634	0.7957	0.8038	**0.8091**
F-score	0.7648	0.7970	0.8048	**0.8102**

Table 6: Performance of integrating discourse classifiers and constraints to polarity classification. Note that the experiments were performed on NTC-8 which contained only opinionated sentences.

(2) *cSSR* enhanced *Baseline* by increasing the average recall by about 15% with only a small drop of precision. The performance of *cSSR* demonstrated that our method could effectively discover high quality common SSRs. The most remarkable improvement was observed on *Continuation* in which the recall increased by almost 20% with only a minor drop of precision. Actually, *cSSR* outperformed *Baseline* in all discourse relations except for *Contrast*. In Discourse Tree Bank (Carlson et al., 2001) only 26% of *Contrast* relations were indicated by cue phrases while in NTC-7 about 70% of *Contrast* were indicated by cue phrases. A possible reason was that we were dealing with Chinese news text which were usually well written. Another important observation was that the performance of *cSSR* was very close to the result of *SVM*.

(3) *SVM+SSRs* achieved the best F-*score* on *Continuation* and average performance. The integration of SSRs to the feature set of *SVM* contributed to a remarkable increase in average F-*score*. The results of *cSSR* and *SVM+SSRs* demonstrated the effectiveness of common SSRs mined by the proposed unsupervised method.

Table 6 presented the performance of integrating discourse classifiers to polarity classification. For *Baseline* and *cSSR*, the information of *nucleus* and *satellite* could be obtained directly from cue-

Relation		*Baseline*	*M&E*	*cSSR*	*SVM*	*SVM* +SSRs
Contrast	**P**	**0.9375**	0.4527	0.7531	**0.9375**	**0.9375**
	R	0.6977	**0.7791**	0.7093	0.6977	0.6977
	F	**0.8000**	0.5726	0.7305	**0.8000**	**0.8000**
Condition	**P**	**1.0000**	0.4444	0.6774	**1.0000**	0.7083
	R	0.5556	**0.8889**	0.7778	0.5185	0.6296
	F	0.7143	0.5926	**0.7241**	0.6829	0.6667
Continuation	**P**	**0.9831**	0.6028	0.9761	0.6507	0.7266
	R	0.2607	0.5865	0.4584	**0.6697**	0.6629
	F	0.4120	0.5945	0.6239	0.6600	**0.6933**
Cause	**P**	**1.0000**	0.5542	0.9429	**1.0000**	0.9412
	R	0.2114	**0.3740**	0.2683	0.2114	0.2602
	F	0.3489	**0.4466**	0.4177	0.3489	0.4076
Purpose	**P**	0.8947	0.3704	0.8163	**0.9167**	0.7193
	R	0.6182	0.7273	0.7273	0.6000	**0.7455**
	F	0.7312	0.4908	**0.7692**	0.7253	0.7321
Average	**P**	**0.9617**	0.5302	0.8864	0.7207	0.7607
	R	0.3410	0.5951	0.4878	0.5856	**0.6046**
	F	0.5035	0.5608	0.6293	0.6461	**0.6737**

Table 5: Performance of recognizing discourse relations. (The evaluation criteria are **P**recision, **R**ecall and **F**-score)

phrase-based patterns and SSRs, respectively. For *SVM+cSSR*, the *nucleus* and *satellite* information was acquired by cSSR if a segment pair could match a cSSR. Otherwise, we used manually annotated *nucleus* and *satellite* information. It's clear that the performance of polarity classification was enhanced with the improvement of discourse relation recognition. *M&E* was not included in this experiment because the performance of polarity classification was decreased by the mis-classified discourse relations. *SVM+SSRs* achieved significant (p<0.01) improvement in polarity classification compared to BPC.

5.4 Discussion

Effect of weighing and filtering

To assess the contribution of weighing and filtering in mining SSRs using a minimum confidence threshold, i.e. $minconf$, we implemented *cSSR'* without weighing and filtering on the same data set. Consider Table 7, *cSSR* achieved obvious improvement in $Precision$ and F-$score$ than *cSSR'*. Moreover, the total number of SSRs was greatly reduced in *cSSR* with only a minor drop of recall. This was because *cSSR'* was affected by thousands of low quality common SSRs which would be filtered in *cSSR*. The result in Table 7 proved that weighing and filtering were essential in our proposed method.

	cSSR'	*cSSR*
Precision	0.6182	**0.8864**
Recall	**0.5014**	0.4878
F-score	0.5537	**0.6293**
NOS	$>$ 1 million	$\approx$ 0.12 million

Table 7: Comparison of *cSSR'* and *cSSR*. "NOS" denoted the number of mined common SSRs.

We further analyzed how the improvement was achieved in *cSSR*. In our experiment, the most common mismatches were auxiliary words, named entities, adjectives or adverbs without sentiments (e.g., "green", "very", etc.), prepositions, numbers and quantifiers. It's straightforward that these words were insignificant in discourse relation classification purpose. Moreover, these words did not belong to the 4 kinds of most representative words. In other words, the weights of most mismatches were calculated using the equation presented in Section 4.2 instead of utilizing a unified value, i.e. -1. Recall Table 3, the weight of "*RB*|*Ka01*" (original: "very") was -0.298 and "*DT*" (original: 'a') was -0.184. Comparing to the weights of mismatches for most representative words (-1.0), the proposed method successfully down weighed the words which were

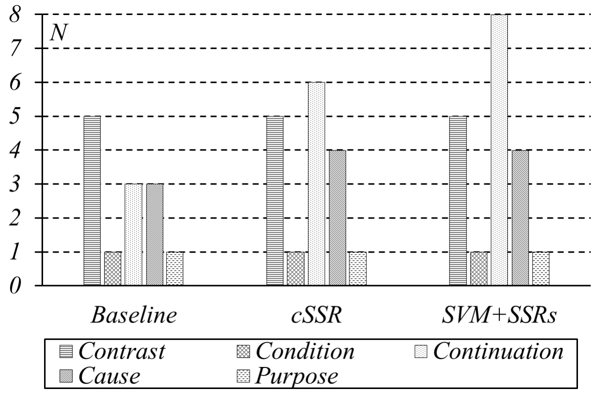

Figure 3: Improvement from individual discourse relations. N denoted the number of ambiguities eliminated.

not important for discourse identification. Therefore, weighing and filtering were able to preserve high quality SSRs while filter out low quality SSRs by setting the confidence threshold, i.e. $minconf$.

Contribution of different discourse relations

We also analyzed the contribution of different discourse relations in eliminating polarity ambiguities. Refer to Figure 3, the improvement of polarity classification mainly came from three discourse relations: *Contrast*, *Continuation* and *Cause*. It was straightforward that *Contrast* relation could eliminate polarity ambiguities because it held between two segments with opposite polarities. The contribution of *Cause* relation also result from two segments holding different polarities such as example (a) in Section 1. However, recall Table 4, although *Cause* occurred more often than *Contrast*, only a part of discourse instances holding *Cause* relation contained two segments with the opposite polarities. Another important relation in eliminating ambiguity was *Continuation*. We investigated sentences with polarities corrected by *Continuation* relation. Most of them fell into two categories: (1) sentences with mistakenly classified sentiments by BPC; (2) sentences with implicit sentiments. For example:

(b) [France and Germany have banned human cloning at present]，[on 20th, U.S. President George W. Bush called for regulations of the same content to Congress] (目前，法国和德国都禁止克隆人的胚胎，美国总统布什20日向国会提出，要求制定同样内容的法规。)

The first segment of example (b) was negative ("banned" expressed a negative sentiment) and a *Continuation* relation held between these two segments. Consequently, the polarity of the second segment should be negative.

6 Conclusions and Future work

This paper focused on unsupervised discovery of intra-sentence discourse relations for sentence level polarity classification. We firstly presented a discourse scheme based on empirical observations. Then, an unsupervised method was proposed starting from a small set of cue-phrase-based patterns to mine high quality common SSRs for each discourse relation. The performance of discourse classification was further improved by employing SSRs as features in supervised methods. Experimental results showed that our methods not only effectively recognized discourse relations but also achieved significant improvement ($p<0.01$) in sentence level polarity classification. Although we were dealing with Chinese text, the proposed unsupervised method could be easily generalized to other languages.

The future work will be focused on (1) integrating more semantic and syntactic information in proposed unsupervised method; (2) extending our method to inter-sentence level and then jointly modeling intra-sentence level and inter-sentence level discourse constraints on polarity to reach a global optimal inference for polarity classification.

Acknowledgments

This work is partially supported by National 863 program of China (Grant No. 2009AA01Z150), the Innovation and Technology Fund of Hong Kong SAR (Project No. GHP/036/09SZ) and 2010/11 CUHK Direct Grants (Project No. EE09743).

References

N. Asher, F. Benamara, and Y.Y. Mathieu. 2008. Distilling opinion in discourse: A preliminary study. *Coling 2008: Companion volume: Posters and Demonstrations*, pages 5--8.

S. Blair-Goldensohn, K.R. McKeown, and O.C. Rambow. 2007. Building and refining rhetorical-semantic relation models. In *Proceedings of NAACL HLT*, pages 428--435.

L. Carlson, D. Marcu, and M.E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*, pages 1--10. Association for Computational Linguistics.

W. Che, Z. Li, and T. Liu. 2010. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13--16. Association for Computational Linguistics.

D.A. Duverle and H. Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 665--673. Association for Computational Linguistics.

C. Fellbaum. 1998. *WordNet: An electronic lexical database*. The MIT press.

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.

W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243--281.

D. Marcu and A. Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368--375. Association for Computational Linguistics.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004. The penn discourse treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Citeseer.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79--86. Association for Computational Linguistics.

L. Polanyi and A. Zaenen. 2006. Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, pages 1--10.

E. Riloff, J. Wiebe, and T. Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25--32. Association for Computational Linguistics.

K. Sadamitsu, S. Sekine, and M. Yamamoto. 2008. Sentiment analysis based on probabilistic models using inter-sentence information.

B. Snyder and R. Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *Proceedings of NAACL HLT*, pages 300--307.

S. Somasundaran, J. Wiebe, and J. Ruppenhofer. 2008. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 801--808. Association for Computational Linguistics.

S. Somasundaran, G. Namata, J. Wiebe, and L. Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170--179. Association for Computational Linguistics.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 149--156. Association for Computational Linguistics.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399--433.

Y.Q. Xia, R.F. Xu, K.F. Wong, and F. Zheng. 2007. The unified collocation framework for opinion mining. In *International Conference on Machine Learning and Cybernetics*, volume 2, pages 844--850. IEEE.

R. Xu and C. Kit. 2010. Incorporating feature-based and similarity-based opinion mining--ctl in ntcir-8 moat. In *Proceedings of the 8th NTCIR Workshop*, pages 276--281.