Chinese Discourse Relation Recognition

Hen-Hsen Huang
Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei, Taiwan
hhhuang@nlg.csie.ntu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei, Taiwan
hhchen@csie.ntu.edu.tw

Abstract

The challenging issues of discourse relation recognition in Chinese are addressed. Due to the lack of Chinese discourse corpora, we construct a moderate corpus with human-annotated discourse relations. Based on the corpus, a statistical classifier is proposed, and various features are explored in the experiments. The experimental results show that our method achieves an accuracy of 88.28% and an F-Score of 63.69% in four-class classification and achieves an F-Score of 93.57% in the best case.

1 Introduction

A discourse relation is the way that two successive arguments logically connected. Recognizing discourse relations attracts much attetions in recent years due to many potential applications. In the annotation scheme of Penn Discourse Treebank 2.0 (PDTB-2.0), the first level of discourse relations includes four classes such as *Temporal*, *Contingency*, *Comparison*, and *Expansion* (Prasad et al., 2008).

The first challenge of discourse relation recognition is the lack of corpus. To construct a discourse corpus has several difficulties. First, the definition of discourse is unclear and varied over different areas. Thus, finding the clear boundary of a discourse argument is a vexing problem by itself. Second, the relationship between arguments is often difficult to decide and inherently subjective. Thus, the annotation and the evaluation are problematic and labor-intensive.

In recent years, the study of discourse relation recognition is growing in the English domain rapidly. One of the reasons is the availability of English corpora with discourse annotations. The two most popular discourse corpora are the Rhetorical Structure Theory Discourse Treebank (RSTDT) (Carlson et al., 2001) and PDTB-2.0. Both of them are based on the Wall Street Journal corpus with human-annotated discourse information. The PDTB-2.0 consists of 36,592 pairs of successive arguments and is tagged with three classes, including *Implicit*, *Explicit*, and *AltLex*, and with the relation types at three levels. Based on these corpora, a number of aspects on discourse relation are explored in these years.

Compared to the English corpora, there is still no Chinese discourse corpus worldwide available. For this reason, the dataset is the first challenge encountered in the study of Chinese discourse relation recognition. To address this issue in this work, we construct a moderate Chinese discourse corpus as a starting point. A supervised statistical classifier is trained and tested on this data set to deal with the problem. Various features are extracted from the corpus and evaluated in the experiments.

The rest of this paper is organized as follows. First, we review the related work in Section 2. In Section 3, the Chinese discourse relation problem is illustrated. Our corpus and the details on the annotation work are presented in Section 4. In Section 5, the method and the features are introduced. The experimental results are discussed in Section 6, and we conclude this paper in Section 7.

2 Related Work

Pitler and Nenkova (2009) reported an explicit discourse relation recognizer that achieves an accuracy of 94.15%. On the other hand, the implicit discourse relation recognition is much more challenging than the explicit one. The implicit discourse relation recognition is to predict the relation of two successive arguments without connectives. In the work of Marcu and Echihabi (2002), the dataset for implicit discourse relation

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1442–1446,
Chiang Mai, Thailand, November 8 – 13, 2011.

detection is automatically derived from explicit samples by removing the connectives. Though this approach is efficient to obtain a large corpus, the pseudo implicit corpus does not exactly capture the property in the real world.

Based on PDTB, in which the argument pairs are distinguished between implicit and explicit, Pitler et al. (2009) and Lin et al. (2009) addressed the task of real implicit discourse relation recognition. In the work of Pitler et al., the implicit discourse relation detection achieves an average accuracy of 62.78% for four-class classification. In the work of Lin et al., the implicit discourse relations are classified into 11 types (selected from the second level tagging in PDTB), and their classifier achieves an accuracy of 40.2%.

From the other aspect, the semi-supervised approach is explored to deal with some relations that are rare in the corpus (Hernault et al., 2010).

3 The Discourse Relation in Chinese

In this work, we adopt the top level classes of PDTB to deal with the Chinese materials.

When two arguments are temporally related, they form a *Temporal* relation. There are two subtypes of Temporal relation, ordered in time (*Asynchronous*, as defined in PDTB-2.0 annotation manual[1]) and overlapped (*Synchronous*). For example, the events in the two arguments in (S1) occur sequentially in time. The event in the second argument happened after the event in the first argument.

(S1) 他首先證實傅爾和中谷義雄的理論。'He first confirmed the theory of Voll and Yoshio Nakatani.'

其次，他發現經絡不僅是電流的良導體，也是電磁波的良導體。 'Second, he found that the meridian is not only a good conductor of current, but also a good conductor of electromagnetic waves.'

The *Contingency* relation talks about the situation that the event in one of the arguments casually affects the event in the other argument. In Chinese, the typical compound connective of Contingency is '因為…，所以…' ('Because…, …'). In sample (S2), the event '颱風來襲' is the cause, and the event '學生停課在家' is its result. Such a relation is defined in PDTB-2.0 as *Cause*, a subtype of Contingency. In sample (S3), the connective '因為…，所以…' is removed. Obviously, the relation between the two clauses is still Contingency. This situation is similar to the case of implicit relation in English.

(S2) 因為颱風來襲，所以學校停止上課。 'Because of the typhoon struck, the school has broken up.'

(S3) 颱風來襲，學校停止上課。 'The typhoon struck; the school has broken up.'

Condition is another typical subtype of Contingency. Condition relation between two arguments specifies the situation in which the event in one argument is conditioned on the event in the other argument.

Comparison is used to show the difference between two arguments. A subtype of Comparison is *Contrast*, where the two arguments share a common predicate or property, and their difference is highlighted.

Expansion, the most common relation, either expands the information for one argument in the other one or continues the narrative flow. In sample (S4), the second argument expands the information to the first argument.

(S4) 伏爾泰是啟蒙運動的領導者，一位偉大的思想家。 'Voltaire is the leader of the Enlightenment, a great thinker.'

除此之外，他也是著作等身的作家。 'In addition, he is also a prolific writer.'

Some words that are usually used as marks to indicate the discourse relations are given in Table 1 for reference.

Relations	Sample Marks
Temporal	同時 (at the same time) 之前 (before)
Contingency	因為 (because) 所以 (therefore) 如果 (if)
Comparison	然而 (however) 雖然 (although) 相反的 (in contrast)
Expansion	而且 (furthermore) 也 (also) 或者 (or) 例如 (for example) 除了 (in addition)

Table 1. Chinese Discourse Relations

[1] http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf

4 Dataset

To deal with the problem of Chinese discourse relation recognition, we firstly construct a corpus for training and testing. The corpus is based on Sinica Treebank 3.1. Total 81 articles are randomly selected from the Sino and Travel sets.

The first issue we encountered is how to segment an article into arguments. In other words, the issue concerns the determination of the argument boundaries. In PDTB, both the boundaries of arguments and the type of discourse relations are annotated by human. In this data set, an argument is not always a sentence. That is, it may be a clause. Sometimes it may be composed of a number of sentences. However, to annotate in such a way is costly and time-consuming. For convenience, we regard an argument as a sentence in this work. A sentence is defined to be a sequence of words ended by a full-stop, a question mark, or an exclamation mark.

In this way, each article is segmented into sentences and shown to three annotators. An annotator assigns one of four discourse relations to each pair of successive sentences. Under this scheme, the annotators regard a sentence as a discourse unit and determine how successive sentences relate to each other. Finally, the majority among the three labels are taken. In the case of ties, an additional annotator will be involved in the final labeling.

The shortage of this annotation scheme is that the samples of Contingency are very rare. In Chinese, the Contingency relation usually occurs inside a sentence. In sample (S3), the two arguments of Contingency, i.e., "因為颱風來襲" and "所以學校停止上課", are two clauses within a single sentence. For this reason, only 94 inter-sentence Contingency relations are tagged in our corpus.

The statistics of the corpus are shown in Table 2. Due to the genre of the Sino and the Travel is descriptive writings, the major relation among the corpus is Expansion.

5 Method

The multi-class support vector machine (SVM) is utilized as our classifier[2]. Due to the unbalance distribution among the four relations, we duplicate the samples of Temporal, Contingency, and Comparison in the training sets proportionally to derive balanced training data.

[2] http://svmlight.joachims.org/svm_multiclass.html

5.1 Features

Length (*Len*): This feature includes the word counts of the first argument, the second argument, the first clause in the first argument, the last clause in the first argument, the first clause in the second argument, and the last clause in the second argument.

***Punctuation* (*Pun*):** The punctuations which end the first and the second arguments are regarded as features. The possible punctuation is a full-stop, a question mark, or an exclamation mark.

***Connective* (*Connect*):** Similar to the connectives in English, some words are usually used as discourse relation marks in Chinese. We prepare a dictionary that contains 319 single words and 489 word pairs. The number of matching words (word-pairs) and their corresponding relation types are considered as features.

***Shared Word* (*SW*):** The number of words shared in the first and the second arguments is considered as a feature. Besides, the common hypernyms shared in both arguments are also counted.

***Word*:** The bags of words in the first argument, in the second argument, and in the first clause of the second argument are considered.

***Part-of-Speech* (*POS*):** The bags of POS in the first argument, in the second argument, and in the first clause of the second argument.

***Hypernym* (*Hyper*):** The bags of hypernym words in the first argument, in the second argument, and in the first clause of the second argument are considered.

***Collocated Word* (*CW*):** Collocated words are the frequent word pairs mined from the training set. The first word and the second word in the pair come from the first argument and the second argument, respectively.

***Number*:** The binary features capture if the dates, the times, the periods, and the numbers exist in the arguments.

6 Experiments

The experimental results for the four relation types are shown in Tables 3, 4, 5, and 6, respectively and the overall performance is given in Table 7. All the performances are evaluated by 5-fold cross-validation.

In general, no single feature is efficient for all the types. For Temporal relation, the feature *Number* contributes the highest recall to capture most candidates. The precision of using single feature only is no more than 25%.

Source	#Articles	#Sentence Pairs	Temporal (#, %)	Contingency (#, %)	Comparison (#, %)	Expansion (#, %)
Sino	27	1594	(104, 6.52)	(51, 3.20)	(156, 9.79)	(1283, 80.49)
Travel	54	1487	(63, 4.24)	(43, 2.89)	(51, 3.43)	(1330, 89.44)
Total	81	3081	(167, 5.42)	(94, 3.05)	(207, 6.72)	(2613, 84.81)

Table 2. Dataset Statistics

Comparatively, the model using all the features achieves a precision of 60.22%. In other words, these features are complementary for recognizing the Temporal relation. The performance is relatively poor for Contingency relation identification. As discussed in Section 4, our annotation does not capture the intra-sentence Contingency relation, thus the most representative examples of Contingency are lost.

The feature *Connect* achieves the highest recall of 70.53% for Comparison relation labeling. The feature *Word*, which achieves a recall of 70.05%, has the similar identification capability. With all the features, an F-Score of 61.24% is achieved.

Expansion is the largest class among the four types. The performance of this type is much better than that of the other three types. Our classifier achieves an F-Score of 93.57% for Expansion.

The performance in macro average is shown in Table 7. Overall, our classifier trained with all features achieves an F-Score of 63.69% and an accuracy of 88.28%.

7 Conclusion

In this work, we address the issue of discourse relation recognition in Chinese. A moderate corpus sampled from Sinica Treebank 3.1 is labeled with discourse relations. The top-level classes used in PDTB are adopted in the data annotation. The SVM classifier trained with various features recognizes the relations between successive arguments automatically. As a result, our classifier achieves an accuracy of 88.28% and an F-Score of 63.69%. In the best case, our classifier achieves an F-Score of 93.57% for the recognition of Expansion relation.

Features	Precision	Recall	F-Score
Len	7.36%	45.51%	12.68%
Pun	5.67%	10.18%	7.28%
Connect	10.07%	41.92%	16.24%
SW	7.54%	23.35%	11.40%
Word	25.23%	64.67%	36.30%
POS	12.76%	65.27%	21.35%
Hyper	13.48%	67.66%	22.49%
CW	25.18%	62.87%	35.96%
Number	7.73%	**73.65%**	13.99%
All	60.22%	67.07%	63.46%

Table 3. Performance of Temporal

The poor performance of the Contingency relation recognition is due to the lack of representative training samples. That needs further investigation.

Features	Precision	Recall	F-Score
Len	3.71%	17.02%	6.10%
Pun	3.07%	20.21%	5.34%
Connect	5.14%	64.89%	9.52%
SW	2.97%	26.60%	5.35%
Word	13.12%	22.34%	16.54%
POS	5.55%	34.04%	9.54%
Hyper	11.81%	37.20%	17.93%
CW	26.09%	44.68%	32.94%
Number	3.28%	15.96%	5.44%
All	50.00%	28.72%	36.49%

Table 4. Performance of Contingency

Features	Precision	Recall	F-score
Len	8.33%	21.74%	12.05%
Pun	6.22%	28.02%	10.18%
Connect	24.79%	**70.53%**	36.68%
SW	7.48%	18.36%	10.63%
Word	34.77%	**70.05%**	46.47%
POS	14.84%	47.83%	22.65%
Hyper	11.81%	37.20%	17.93%
CW	24.29%	62.32%	34.96%
Number	10.83%	24.64%	15.04%
All	60.66%	61.84%	61.24%

Table 5. Performance of Comparison

Features	Precision	Recall	F-score
Len	85.90%	35.44%	50.18%
Pun	84.32%	39.72%	54.01%
Connect	91.48%	21.35%	34.63%
SW	85.84%	39.92%	54.49%
Word	93.35%	**74.17%**	82.66%
POS	94.00%	35.36%	51.39%
Hyper	92.96%	30.81%	46.28%
CW	96.10%	72.52%	82.66%
Number	90.75%	19.52%	32.13%
All	93.27%	93.88%	93.57%

Table 6. Performance of Expansion

Features	Precison	Recall	F-Score	Accuracy
Len	26.33%	29.93%	20.25%	34.50%
Pun	24.82%	24.53%	19.20%	36.74%
Connect	32.87%	49.67%	24.27%	27.10%
SW	25.96%	27.06%	20.47%	37.16%
Word	41.62%	57.81%	45.49%	71.79%
POS	31.79%	45.62%	26.23%	37.78%
Hyper	30.84%	43.76%	23.93%	33.50%
CW	42.91%	60.60%	46.63%	70.46%
Number	28.15%	33.44%	16.65%	22.69%
All	66.04%	62.88%	63.69%	88.28%

Table 7. Overall Performance

References

L. Carlson, D. Marcu, and M. E. Okurowski. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of Second SIGDIAL Workshop on Discourse and Dialogue-Volume 16*, pages 1-10.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A Semi-Supervised Approach to Improve Classification of Infrequent Discourse Relations using Feature Vector Extension. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 399-409.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343-351.

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 (ACL 2002)*, pages 368–375, Morristown, NJ, USA.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic Sense Prediction for Implicit Discourse Relations in Text. In *Proceedings of the 47th annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 683-691.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of the 47th annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*. Short Papers.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC'08*.