Cross-lingual Identification of Ambiguous Discourse Connectives for Resource-Poor Language

Lanjun Zhou[1] *Wei Gao*[2] *Binyang Li*[1] *Zhongyu Wei*[1] *Kam-Fai Wong*[1]
[1] Dept. of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
[2] Qatar Computing Research Institute, Qatar Foundation
{ljzhou,byli,zywei,kfwong}@se.cuhk.edu.hk, wgao@qf.org.qa

Abstract

The lack of annotated corpora brings limitations in research of discourse classification for many languages. In this paper, we present the first effort towards recognizing ambiguities of discourse connectives, which is fundamental to discourse classification for resource-poor language such as Chinese. A language independent framework is proposed utilizing bilingual dictionaries, Penn Discourse Treebank and parallel data between English and Chinese. We start from translating the English connectives to Chinese using a bi-lingual dictionary. Then, the ambiguities in terms of senses a connective may signal are estimated based on the ambiguities of English connectives and word alignment information. Finally, the ambiguity between discourse usage and non-discourse usage were disambiguated using the co-training algorithm. Experimental results showed the proposed method not only built a high quality connective lexicon for Chinese but also achieved a high performance in recognizing the ambiguities. We also present a discourse corpus for Chinese which will soon become the first Chinese discourse corpus publicly available.

Keywords: Discourse, Explicit Connectives, Ambiguity of Connectives.

Proceedings of COLING 2012: Posters, pages 1409–1418,
COLING 2012, Mumbai, December 2012.

1 Introduction

Discourse classification with its applications in many natural language processing tasks such as automatic summarization (Spärck Jones, 2007), text generation (McKeown, 1992) and sentiment analysis (Zhou et al., 2011) etc., has attracted much attention in recent years. However, the lack of annotated corpora brings limitations in research of discourse for many languages (e.g., Chinese).

The Penn Discourse Tree Bank 2.0 (PDTB2) (Prasad et al., 2008a) divided the English discourse connectives into two categories: explicit connectives and implicit connectives. Explicit discourse connectives could be found within a sentence or between sentence pairs while implicit connectives appear only between paragraph-internal adjacent sentence pairs. We focus on the explicit connectives in this work.

Pitler et al. (2008) argued that discourse senses triggered by explicit connectives were easy to be identified in English PDTB2. However, their conclusions may not be true for other languages (Alsaif and Markert, 2011). The ambiguities of explicit connectives could vary among different languages. For many other languages (e.g., Chinese), there is no published discourse corpus available rendering even the identification of explicit discourse difficult. In this work, we focus on the problem of identifying explicit discourse connectives and recognizing their ambiguities for languages without annotated corpus, where the problem is dealt with from cross-lingual perspective. We set English as the source language and Chinese as the target language and we attempt to get answers to the following two questions: (1) Is it possible to build a high quality discourse connective lexicon for the target language (i.e., Chinese) from the source language (i.e., English)? (2) How to disambiguate the ambiguities of each discourse connective in the target language, including the ambiguities between discourse usage to non-discourse usage (e.g., 'for' serves both discourse function and non-discourse function) and ambiguities among the discourse relations it may signal (e.g., 'since' could signal both *Temporal* and *Causal* relations)? To the best of our knowledge, our work is the first that addresses the identification of two different kinds of ambiguities for resource-poor language.

To answer the above questions, we propose a language independent framework using bilingual dictionaries, annotated corpora from the source language and parallel data. The framework mainly consists of the following steps: (1) translate the English connectives to Chinese using a bi-lingual dictionary and expand the connective set by adding synonyms; (2) extract all English connectives aligned with each of the Chinese connectives in large amount of bilingual parallel data; (3) recognize and disambiguate the ambiguities of Chinese connectives. The experimental results showed the effectiveness of the proposed method.

We also present the Discourse Treebank for Chinese (DTBC) project as there is no published discourse corpus in Chinese. Currently, DTBC contains discourse annotations for 500 articles selected from the Penn Chinese Tree Bank 6 (Xue et al., 2005). We annotated 2,549 explicit relations with connectives and arguments. DTBC will soon become the first Chinese discourse corpus publicly available.

2 Related Work

Ambiguity of Connectives. Pitler et al. (2008) argued that the overall degree of ambiguity for English connectives were low. Alsaif and Markert (2011) showed that Arabic connec-

tives are more ambiguous. The most closely related work was Versley (2010). They treated the two kinds of connective ambiguities without necessary differentiation. However, these two kinds of ambiguities should be studied individually since they are essentially different (Pitler and Nenkova, 2009b). Therefore, the way we dealt with ambiguities was very different from theirs. To the best of our knowledge, there is little work that focuses on the problem of cross-lingual identification of ambiguities of discourse connectives for discourse classification.

Discourse corpus annotation. For English, there are mainly two corpora: (1) RST Discourse Treebank (RST-DT) (Carlson et al., 2001) following the RST (Mann and Thompson, 1988); (2) Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004)(Prasad et al., 2008a). Based on the RST or PDTB, corpora for other languages such as Spanish (da Cunha et al., 2011), Hindi (Prasad et al., 2008b), Arabic (Al-Saif and Markert, 2010) etc. were developed. However, for most of the other languages, there is no published discourse corpus.

For Chinese, Xue (2005) proposed the Chinese Discourse Treebank (CDTB) Project. However, they mainly discussed the issues that arise from the annotation process and the annotated corpus was not published. Zhou et al. (2011) annotated 1,225 intra-sentence discourse instances for improving the performance of polarity classification for Chinese. However, the discourse scheme proposed by them was specially for sentiment analysis. Zhou and Xue (2012) presented a PDTB-style discourse corpus for Chinese. Nevertheless, their data was not publicly available. As far as we know, there is no published discourse corpus in Chinese.

3 Methods

3.1 Finding possible discourse connectives

Utilizing the most frequent discourse relation a connective may signal, Pitler et al. (2008) achieved over 90% of accuracy in recognizing explicit relations in PDTB2 and Alsaif and Markert (2011) reported 82.7% of accuracy in Arabic. As a result, building a high quality connective lexicon is crucial for recognizing explicit relations in the target language.

Since English explicit connectives could be extracted directly from PDTB2, the most intuitive way of finding discourse connectives in the target language is the dictionary based approach. Thus, we adopt an English-Chinese bilingual dictionary[1]. Similar resources could be found between other language pairs. We first extract the Chinese translations for all English connectives using the bilingual dictionary. Then, the connectives are extended using the Chinese synonym list extended version (Che et al., 2010). Note that we adopt part-of-speech restrictions according to the settings of PDTB2 during the translation and extension process. However, many of the connectives in the list are noisy or ambiguous (See section 4). Hence, the connective list need to be refined to preserve only high quality connectives.

3.2 Filtering and estimating the ambiguities of discourse connectives

During the translation process of Section 3.1, we found that the ambiguity of a connective in English could usually be eliminated when translated to Chinese. For example, 'since' could be translated to unambiguous Chinese connectives signaling different discourse relations (e.g., (因为, *Contingency*), (自从, *Temporal*)). Although this observation is between

[1]The 21st Century Unabridged English-Chinese Dictionary

English and Chinese, we believe that similar findings also occur between other language pairs. This observation inspired us to use word alignment information for estimating the ambiguities of Chinese discourse connectives. Fortunately, there are large amount of parallel data available between Chinese and English.

The general idea of the proposed method is to estimate the ambiguities of Chinese connectives by calculating the entropy over its probability distribution on parallel data. Suppose S denotes the source language, T denotes the target language, $E=\{e_1, e_2...e_n\}$ denotes all discourse connectives in T, $E'=\{e'_1, e'_2...e'_m\}$ denotes all discourse connectives in S and $R=\{r_1, r_2, r_3, r_4\}$ denotes the top level relation (i.e., *Temporal*, *Contingency*, *Comparison* and *Expansion*) in PDTB2. For $e'_i \in E'$ and $r_k \in R$, we estimate $P(r_k|(e'_i, S))$ using the distribution of occurrences for e'_i over R.

Given a discourse connective $e_j \in E$ from the target language, suppose $C_j=\{c_1, c_2...c_m\}$ denotes the frequency of occurrences for each connective in E' aligned with e_j in the parallel data. The probability for e_j signals relation r_k in T is estimated using the following equation:

$$P(r_k|(e_j, T)) = \sum_i P(r_k|(e'_i, S)) \frac{c_i}{||C_j||} \quad (1)$$

in which $||C_j|| = \sum c_i$. As entropy is a measure of uncertainty, the ambiguity of e_j in T is estimated using the following equation:

$$Amb(e_j, T) = -\sum_k P(r_k|(e_j, T)) \cdot \log P(r_k|(e_j, T)) \quad (2)$$

Finally, we rank all connectives in E and use a threshold value $max\text{-}e$ to control the quality of E. $max\text{-}e$ will be determined experimentally to achieve the best performance in recognizing the explicit relations in Chinese.

3.3 Identifying discourse usage of connectives

Given a discourse connective, it could be ambiguous between discourse usage and non-discourse usage in different contexts. For example, in most of the cases, the English connective 'for' does not act as a discourse connective. Since we assume that there is no annotated corpus for the target language, we adopt the co-training algorithm(McKeown, 1992). The main idea of our method is to start from annotated data in English (i.e., the source distribution) and then increase the size of training data by incrementally adding the unlabeled data from Chinese (i.e., the target distribution). We outline the steps of proposed co-training based method in Algorithm 1.

Note that all labeled and unlabeled instances will have two versions: an English version and a Chinese version. We adopt the Baidu Translator[2] for the translation process between English and Chinese. c_1 and c_2 will output probabilities of every testing instance for whether it will serve a discourse function. We take the average of the probabilities given by c_1 and c_2 as the prediction of Algorithm 1.

Since the performance of discourse vs non-discourse usage classification reported by Pitler and Nenkova (2009a) had already reached near human results for English, we adopt their

[2]http://translate.baidu.com/

Algorithm 1 Co-training algorithm for identifying discourse usage of connectives

Given:
- a feature set F_e for the English view
- a feature set F_c for the Chinese view
- a set L of labeled instances from PDTB2;
- a set U of unlabeled instances from DTBC;

Create a pool U' of examples by choosing u examples randomly from U
Loop for k iterations:
  Use L to train a classifier c_1 that uses the feature set F_e
  Use L to train a classifier c_2 that uses the feature set F_c
  Allow c_1 to label U' and choose p most-confident positive instances and n negative instances
  Allow c_2 to label U' and choose p most-confident positive instances and n negative instances
  Add these self-labeled examples to L
  Replenish U' by randomly choose $2*(p+n)$ from U

	Temporal	*Contingency*	*Comparison*	*Expansion*
DTBC	10%	17%	13%	61%
PDTB2	19%	19%	29%	33%

Table 1: Distribution of explicit relations in DTBC and PDTB2.

feature set for the English view. The Chinese view comprises syntactic features, lexical features and word alignment features. Lexical features and syntactic features are inspired by previous work (Pitler and Nenkova, 2009a; Alsaif and Markert, 2011). The word alignment features are new. Intuitively, given a sentence (or sentence pair) from the source language, if a connective signals a discourse relation, the translation of this connective (if any) will signal the same relation in the target language. Hence, word alignment information will be useful for recognizing discourse usage for the target languages.

4 Experiments and Discussion

4.1 Data

PDTB2. We utilized the Penn Discourse Treebank 2 (PDTB2) (Prasad et al., 2008a), the largest annotated corpora available for English.

DTBC: We presented the Discourse Treebank for Chinese (DTBC). DTBC followed the observations of CDTB (Xue, 2005) and principles of PDTB2 as far as possible. At the current stage, we only annotated explicit discourse relations with their corresponding connective and arguments. DTBC consists of discourse annotations for 500 Chinese news texts selected from Penn Chinese Tree Bank 6 (CTB6) (Xue et al., 2005). It contains annotations for 2,549 explicit relations with connectives and arguments. 2 human annotators were trained to annotate discourse information for all articles. The *kappa-value* is $k_e = 0.78$ for relation identification for the top-level relations. A statistics of DTBC is shown in Table 1. We adopt this corpus to evaluate the performance of discourse usage vs non-discourse usage and explicit discourse relation classification.

NiuTrans: An open-source English-Chinese statistical machine translation system[3]. It contains a sample data of 199,630 English-Chinese parallel sentences. The word alignment

[3]http://www.nlplab.com/NiuPlan/NiuTrans.html

results were the output of GIZA++ (Och and Ney, 2003).

4.2 Experimental settings

4.2.1 Building discourse connective lexicons for Chinese

DIC-1 & DIC-2: The method described in Section 3.1. If a Chinese connective was translations of multiple English connectives, we chose the English connective appeared most frequently in PDTB2 for DIC-1 while the connective will be removed in DIC-2.

DIC+ENT: Different with DIC-2, we did not drop the ambiguous connectives. Instead, the ambiguities in terms of different relations a connective may signal were estimated using the method proposed in Section 3.2. Note that we only estimated the ambiguities in this paper, disambiguating the ambiguities would be another work.

DIC-1, DIC-2 and DIC+ENT output three different discourse connective lexicons. Then, three annotators were trained to label all the connectives as 'discourse connective' or 'not discourse connective'.The golden set was built according to the majority voting.

It was also interesting to evaluate the performance of discourse classification using the lexicons generated by above methods. Note that in this experiment, we used annotated discourse usage information for all connectives in DTBC. A connective based classifier (Pitler et al., 2008) was utilized to evaluate the performance of discourse classification for Chinese. Moreover, we compared the performance of the above methods to the following machine translation based method.

MT-1: We adopted the Baidu Translator[4] to translate all the Chinese text to English. Then, we find discourse relations in the translated English texts.

4.2.2 Identifying discourse usage of connectives

In this experiment, we utilized all sentences containing connectives in DTBC. Since many of the sentences contained more than one discourse connective, the annotated connectives were added to the positive set and others were added to the negative set. We adopted a maximum entropy classifier[5] with iteration number of $i = 15$. We empirically set $|U'| = |U| = u$, $p = 5$, $n = 5$ for the co-training based methods.

CON : A connective would serve a discourse function when it appeared.

MT-2 : We implemented the state-of-the-art method proposed by Pitler and Nenkova (2009b). We adopted PDTB2 as the training data and English translation of DTBC as the testing data.

COT-1 & COT-2 : The method described in Section 3.3. In COT-1, we adopted the same feature set for English and Chinese. The feature set included connectives and syntactic information. The Stanford Parser[6] was adopted to get the syntax structures for translated English sentences and all Chinese sentences. COT-2 added lexical features and alignment features to the Chinese view.

[4] http://translate.baidu.com/
[5] http://mallet.cs.umass.edu
[6] http://nlp.stanford.edu/software/lex-parser.shtml

	Size	*Precision*	*Recall*	*F-score*
DIC-1	561	0.5009	**1.0000**	0.6675
DIC-2	413	0.4649	0.6833	0.5533
DIC+ENT	231	**0.8615**	0.7082	**0.7773**

Table 2: Performance of different connective lexicons. Note that we set $max\text{-}e$=1 for DIC+ENT because we did not need to drop any ambiguous connectives in this experiment

	DIC-1	**DIC-2**	**MT-1**	**DIC+ENT**
Precision	0.7948	**0.9253**	0.9082	0.8119
Recall	0.5982	0.3967	0.4287	**0.6937**
F-score	0.6827	0.5554	0.5825	**0.7481**

Table 3: Performance of discourse classification on DTBC. The result of DIC+ENT was acquired by setting $max\text{-}e = 0.3$.

4.3 Results

4.3.1 Results of building discourse connective lexicons

Refer to Table 2, DIC+ENT significantly outperformed DIC-1 and DIC-2 in both *precision* and *F-score*. Although the recall of DIC+ENT was not high, the most common discourse connectives in Chinese were all recognized (Refer to Table 3). We believed that the *recall* of DIC+ENT will be further improved by adding more parallel data. Moreover, the size of connective lexicon was greatly reduced in DIC+ENT. Noticeably, the performance of DIC-2 was poor comparing to DIC-1. The recall of DIC-2 dropped to 0.6833 since we filtered 148 connectives which were ambiguous. The result of DIC-2 indicated that over 30% of the Chinese connectives were ambiguous. Accordingly, it was important to recognize the ambiguity of each connective before the discourse classification task.

4.3.2 Results of discourse classification

We introduced $max\text{-}e$ to control the quality of discourse connectives in DIC+ENT. If $max\text{-}e$= 0, the proposed method became DIC-2. This threshold was tuned using the development data (20% of DTBC). The best performance was observed when $max\text{-}e$=0.3 ($F\text{-}score$= 0.7481). Accordingly, we adopted this optimal value for $max\text{-}e$ in the following experiment.

Table 3 shows the experimental results of explicit relation classification. Consider Table 3, following conclusions could be drawn: (1) DIC+ENT reached the best result for *recall* and *F-score*. Note that the performance of DIC+ENT outperformed DIC-1 in both *precision* and *recall* . This observation indicated the effectiveness of proposed method. (2) The comparison between DIC-2 and DIC+ENT showed that the drop of recall for DIC-2 comparing with DIC+ENT is up to 0.26. This indicated that ambiguous connectives cannot to be neglected in explicit relation classification. (3) The performance of MT-1 was poor comparing to DIC+ENT. The reason mainly lies in two aspects: (a) the machine translation results were far from perfect; (b) the PDTB2 only contained annotations for 100 different English connectives, resulting to a low recall.

	CON	MT-2	COT-1 ($k = 41$)	COT-2 ($k = 34$)	PDTB2* (reported*)
Accuracy	0.2470	0.6404	**0.7043**	0.7428	0.8586
F-score	0.3961	0.6814	**0.7590**	0.7933	0.7533

Table 4: Experimental results of discourse usage identification for Chinese. The best results of COT-1 and COT-2 during the iteration were presented in the table. *The result was reported by (Pitler and Nenkova, 2009b) on PDTB2.

4.3.3 Results of identifying discourse usage of connectives

Table 4 presents the results of discourse usage identification for Chinese. Refer to Table 4, the co-training based methods significantly ($p < 0.05$) outperformed CON and MT-2 in identification of discourse usages for Chinese connectives.

The performance of CON which predicted discourse usage for every occurrence of connective was poor. However, Pitler and Nenkova (2009b) reported 85.86% of *accuracy* and 75.33% of *F-score* for the connective only method on English PDTB2. We performed a simple error analysis for CON and found that some Chinese connective served as non-discourse function appeared very frequently in DTBC. For example, '和 (and)' and '在 (in, at, etc.)' appeared thousands of times in DTBC but served as discourse function less than 5% of the time. Thus, the disambiguation between discourse usage to non-discourse usage in Chinese DTBC was essential and more challenging than in English.

MT-2 performed better than the connective only method. However, it only achieved less than 50% of *recall* since the English translations of some common Chinese connectives not belonged to the PDTB2 connective list. Moreover, the parsing results were inaccurate because of the imperfect translations and long sentences. Thus, the overall performance of MT-2 was not satisfactory.

The comparison between COT-1 and COT-2 showed that lexical features and alignment features were effective. One possible explanation was that the performance of proposed method highly relied on the results of machine translation and syntactic parsers. The lexical features and alignment features could still provide useful information when accurate machine translation or syntactic information were unavailable.

Conclusion and perspectives

In this paper, we proposed a language independent framework for building discourse connective lexicons and recognizing their ambiguities for languages without annotated corpora; Experimental results showed the effectiveness of our method. The future work includes: (1) Adapt the proposed method to other languages such as Arabic, Hindi, etc; (2) continue the annotation work of DTBC to include journal articles and well written reviews.

Acknowledgments

This work is partially supported by National 863 program of China (No. 2009AA01Z150), Innovation and Technology Fund of Hong Kong (No. InP/255/10, GHP/036/09SZ) and CUHK Direct Grants (No. 2050525).

References

Al-Saif, A. and Markert, K. (2010). The leeds arabic discourse treebank: Annotating discourse connectives for arabic. In *Language Resources and Evaluation Conference (LREC)*.

Alsaif, A. and Markert, K. (2011). Modelling discourse relations for arabic. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 736--747.

Carlson, L., Marcu, D., and Okurowski, M. (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.

Che, W., Li, Z., and Liu, T. (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.

da Cunha, I., Torres-Moreno, J., and Sierra, G. (2011). On the development of the rst spanish treebank. *ACL HLT 2011*, page 1.

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

McKeown, K. (1992). *Text generation: using discourse strategies and focus constraints to generate natural language text*. Cambridge Univ Pr.

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

Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19--51.

Pitler, E. and Nenkova, A. (2009a). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13--16, Suntec, Singapore. Association for Computational Linguistics.

Pitler, E. and Nenkova, A. (2009b). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13--16. Association for Computational Linguistics.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. (2008). Easily identifiable discourse relations. *Proceedings of COLING 2008, Posters Proceedings*, pages 87--90.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008a). The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961--2968. Citeseer.

Prasad, R., Husain, S., Sharma, D., and Joshi, A. (2008b). Towards an annotated corpus of discourse relations in hindi. *Proceedings of IJCNLP-2008*.

Spärck Jones, K. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449--1481.

Versley, Y. (2010). Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83--82.

Xue, N. (2005). Annotating discourse connectives in the chinese treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 84--91. Association for Computational Linguistics.

Xue, N., Xia, F., Chiou, F., and Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(02):207--238.

Zhou, L., Li, B., Gao, W., Wei, Z., and Wong, K. (2011). Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the 2011 Conference on Empirical methods in natural language processing*, pages 162--171. Association for Computational Linguistics.

Zhou, Y. and Xue, N. (2012). Pdtb-style discourse annotation of chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69--77, Jeju Island, Korea. Association for Computational Linguistics.