Towards a Discourse Relation-aware Approach for Chinese-English Machine Translation

Frances Yung
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192 Japan
pikyufrances-y@is.naist.jp

Abstract

Translation of discourse relations is one of the recent efforts of incorporating discourse information to statistical machine translation (SMT). While existing works focus on disambiguation of ambiguous discourse connectives, or transformation of discourse trees, only explicit discourse relations are tackled. A greater challenge exists in machine translation of Chinese, since implicit discourse relations are abundant and occur both inside and outside a sentence. This thesis proposal describes ongoing work on bilingual discourse annotation and plans towards incorporating discourse relation knowledge to a Chinese-English SMT system with consideration of implicit discourse relations. The final goal is a discourse-unit-based translation model unbounded by the traditional assumption of sentence-to-sentence translation.

1 Introduction

Human translation is created at document level, suggesting that translation of a particular sentence depends also on the 'discourse structure'. Recently, some MT researchers have started to explore the possibility to incorporate linguistic information outside the sentence boundary for MT, such as topical structure, coreference chains, and lexical coherence. Among various discourse structures, discourse relations, also known as coherence relations, are meaningful relations connecting text segments and are crucial to the human cognitive processing as well as memory of texts (Sanders and Noordman, 2000). These relations can be explicitly marked in a text by signaling phrases or implicitly implied. Even when they are explicit, some markers are ambiguous and do not always signal the same relation. In addition, strategies to represent discourse relations vary across languages. It is thus a challenging task to correctly translate discourse relations.

This thesis proposal presents my plan towards building a discourse-relation-aware machine translation system translating from Chinese to English. In particular, I would like to focus on modeling the translation of implicit discourse relations, which has not yet been exploited to date to my knowledge, but is yet a noticeable problem since implicit discourse relations are abundant in Chinese. According to the statistics of the bilingual discourse annotation in progress, about 1/4 of the Chinese implicit DCs are translated to explicit DCs in English.

A reasonable initial attempt to learn discourse-relation-aware translation rules is a knowledge-based approach based on an annotated corpus. This proposal describes my ongoing work on annotating and cross-lingually aligning discourse relations in a Chinese-English translation corpus, as well as my plans to incorporate the resulting linguistic markup into an SMT system. Motivated by the characteristics of long Chinese sentences with multiple discourse segments, a further direction of the research is to translate in units of discourse segments instead of sentences.

Section 2 gives an overview of existing literature. Section 3 explains the motivations behind my research on discourse relations for MT. Section 4 describes my ongoing work of bilingual discourse annotation, followed by statistics to date . Section 5 present my plans for next steps. Finally, a conclusion is drawn in Section 6.

2 Survey

2.1 English discourse processing

There are a number of discourse-annotated English resources, including the 'RST Treebank' (Carlson et al., 2001) and the 'Discourse GraphBank' (Wolf and Gibson, 2005), which consist

Proceedings of the ACL 2014 Student Research Workshop, pages 18–25,
Baltimore, Maryland USA, June 22-27 2014.

of 385 and 135 articles respectively. Recent discourse research often make use of the large-scaled Penn Discourse Treebank (PDTB) (Prasad et al., 2008). Departed from annotation using predefined discourse relations, such as 'Rhetorical Structure Theory' (Mann and Thompson, 1988), PDTB introduces a lexically-ground formalism to annotate discourse relations by identifying the discourse connectives (DCs). An example is shown in the following.

Example 1: Since **McDonald's menu prices rose this year**, *the actual decline may have been more.* (PDTB 1280)

'Since' is an explicit DC taking the *italic segment* as the first argument (Arg1), and the **bolded segment** as the second argument (Arg2), which is syntactically attached to the DC. Implicit DCs are inserted by annotators between adjacent sentences of the same paragraph to represent inferred discourse relations. Each DC is annotated with defined *senses* classified into 3 levels of granularity.

PDTB allows evaluation of English discourse parsing tasks and disambiguation tasks (Pitler and Nenkova, 2009; Lin et al., 2010), which reveal that implicit discourse relations are much harder to learn than explicit discourse relations (Pitler et al., 2009; Zhou et al., 2010). For example, classification of the 4 main relation senses (temporal, contingency, comparison, expansion) reaches 94% accuracy for explicit relations (Pitler and Nenkova, 2009), but only range from F-scores of 20% for 'temporal' to 76% for 'expansion' relations, possibly due to unbalanced number of training instances (Pitler et al., 2009; Zhou et al., 2010).

2.2 Chinese discourse processing

Schemes for Chinese discourse annotation have been proposed in the existing literature (Xue, 2005; Zhou and Xue, 2012) but the corresponding resource is not yet available. Zhou et al. (2012) proposed to project English discourse annotation and classification algorithms to Chinese data, but the transfer was based on automatic word alignment and machine translation results. Works in Chinese discourse parsing report F-scores of 64% in classification of inter-sentence discourse relations and 71% in 2-way classification of intra-sentence contingency and comparison relations (Huang and Chen, 2011; Huang and Chen, 2012), training on a moderately sized (81 articles) corpus and considering explicit and implicit relations collectively. Corelation between discourse relation and sentiment was also explored based on annotated data (Huang et al., 2013).

2.3 Discourse relations in SMT

Earlier studies of discourse relations in MT includes Marcu et al. (2000), which proposed a discourse transfer model to re-construct the target discourse tree from the source discourse tree, parsed by the (RST). However, incorporation to an SMT system was not discussed in the work. Recent works focus on the translation of ambiguous DCs, such as 'since' in the temporal sense vs. 'since' in the reason sense. This is achieved by annotating the DCs in the training data by 'translation spotting', which is to manually align the DCs of the source text to their translation in the target text, either occurring as DCs or other expressions (Meyer et al., 2011; Popescu-Belis et al., 2012; Meyer et al., 2012; Meyer and Polakova, 2013; Cartoni et al., 2013). Experiments of these works have been conducted in English-to-French, Czech and German translation and only explicit DCs were considered.

Tu et al. (2013) proposed a framework for Chinese-to-English translation, in which the source text is automatically parsed by an RST parser and translation rules are extracted from the source discourse trees aligned with the target strings. An improvement of 1.16 BLEU point is reported, considering only intra-sentential explicit relations.

Meyer et al. (2012) found that the translation of DC improves by up to 10% disregarded of BLEU, which stays around the baseline system score. To detect the improvement, they used a metric known as *ACT* (Accuracy of Connective Translation) (Hajlaoui and Popescu-Belis, 2012; Hajlaoui and Popescu-Belis, 2013), which relies on bilingual word alignment and a dictionary of DCs. In the setting, missing/additional DC (i.e. potential implicitation/explicitation of discourse relations) are to be checked manually for the validity.

3 Motivation

The motivation behind a discourse-relation-aware translation model for Chinese is two-fold. First of all, on top of ambiguous discourse connectives as in other languages, Chinese documents contain

abundant implicit connectives (Xue, 2005). In particular, complex sentences often occur in the form of 'running sentences', in which loose clauses run in a sequence separated by commas yet without explicit connectives. Such sentence structures are used to represent the temporal or reasoning order or related events, or simply to achieve consistent rhythmic patterns. In contrast, syntactical constraint is prominent in English and this kind of 'paratactic' structures only occur as occasional rhetorical measures. In other cases, relations between clauses within a sentence are marked by coordinating or subordinating conjunctions in order to maintain an intact sentence structure.

Another motivation is that translation in units of sentences is not always preferable in Chinese-English translation. In fact, each comma-separated segment of a 'running sentence' can be considered as an elementary discourse units (EDU) (Yang and Xue, 2012; Zhou and Xue, 2012) and aligned across the two languages. In current SMT models, sentence splitting is the result of the language model or translation rules containing periods or sentence initial markers. A long Chinese 'running sentence' is typically translated to one English sentence with 'comma splices' (ungrammatical commas between complete sentences without connecting by conjunctions). On the other hand, discourse structure provides clues to split the source sentence. It is because some DCs only relate EDUs within the same sentences (e.g. *'but', 'because'*) while some only relate with the previous sentence (e.g. *'however', 'in addition'*)(Stepanov and Riccardi, 2013).

Example 2 shows two versions of English translation of a Chinese sentence as output by *Google Translate*. Note that in the original Chinese sentence, all the DCs are omitted to achieve a quadruplet pattern. Implicit DCs, represented by glossed words in brackets, can be inserted to each comma-separated clause to signal the discourse relations. Without explicit DCs, the MT output **(MT original)** results in a sequence of broken clauses, whereas with inserted DCs **(MT w/DC)**, the clauses are joined by the translated DCs to a complete sentence. In addition, the dropped pronoun 'you' is properly generated, potentially due to improvement in syntactical parsing of the source sentence.

Example 2

Source: (如果-if)交納税款有困難的，(便-then)可暫緩積欠，(但是-but)新税不欠，(而且-furthermore)掛税免罰，(並-and)逐年繳清。

MT original: Difficult to pay taxes, may suspend arrears, the new tex is not owed, penalties linked tax free, paid annually.

MT w/DC: **If** you have difficulty to pay taxes, you can suspend the arrears, **but** the new tax is not owed **and** taxes linked to impunity **and** paid annually.

Ref: Those having difficulty paying taxes can temporarily postponing old debt **but** not owing on new taxes, **and** suspending taxes **and** waiving fines, **and** paying off year by year.
(adapted from Chinese Tree Bank Art.89)

4 Work in progress: Cross-lingual annotation of discourse relations

Towards building a statistical machine translation system that tackles discourse relations specifically, I started manually annotating a Chinese-English translation corpus with discourse relations. The purpose of annotation is not only to create data but also to understand the problems in Chinese discourse processing and translation. The completed annotation is planned to be released.

Comparing with representation of discourse relations by analytical definitions, the PDTB-styled association of discourse relations to lexical connectives is more compatible to the procedures of statistical machine translation. Therefore, the PDTB convention is adopted for the annotation of connectives on both sides of the parallel corpus. Instead of sense annotation, the DCs are aligned in similar manner as the 'translation spotting' approach (Meyer et al., 2011; Popescu-Belis et al., 2012; Cartoni et al., 2013). In other words, the 'senses' are disambiguated by the translation of the DCs. The data used is the English Chinese Translation Treebank (Bies et al., 2007), which consists of 325 Chinese news stories translated into 146,300 words of English. Adaptations made to capture the cross-lingual difference in discourse relations are explained in the following.

4.1 EDU segmented by punctuations

In the PDTB, the span of each EDU (Arg1 or Arg2), which can range from a single noun to multiple sentences, are manually annotated. While

each WSJ paragraph[1] contains three sentences on average, the typical 'running sentences' in Chinese are exceptionally long. It is hard for annotators to agree on an EDU span, and neither does it have direct effect on the DC translation. Therefore, I follow previous works (Yang and Xue, 2012; Zhou and Xue, 2012) and consider a segment separated by Chinese punctuations, especially commas, as the span of an EDU.

Nonetheless, there are exceptions since Chinese commas are used arbitrarily to signify '*pauses*' in the sentence. Three original tags are defined to annotate the exceptions: '**ATT**ribution' , 'initialized **ADV**erbial', and '**OPT**ional comma' (refer to Table 1). These are designed for training of automatic EDU segmentation.

4.2 Explicit DCs

After recognizing a valid EDU on the source text, explicit DC(s) in the EDU are tagged '**EXP**' and aligned to their translation on the target side, which are not necessarily explicit DCs. In contrast with the defined list of subordinating conjunctions, coordinating conjunctions and adverbials, DCs are not limited to any syntactical categories in this scheme so as to improve the coverage of cross-lingual annotation. For example, 'at the same time' and 'in spite of the fact that' are annotated as DC instances, since they function as the DCs 'simultaneously' and 'although' respectively, independent of context.

In addition, conjunctions between VP constructions, which are not annotated in the PDTB, are also annotated as explicit DCs. It is because subjects are often dropped in Chinese and many EDUs will be ignored if VP constructions are excluded.

4.3 Discourse markers alternative to DCs

Discourse relations can be explicitly marked by non-DC expressions that are context dependent. Following the PDTB scheme, the '**ALT**Lex' tag is used to annotate such alternative lexicalization of discourse relations. However, with a loose definition of DC, few alternative expressions are identified. Therefore, the '**ALT**' tag is defined only on the English side, which particularly serves to mark non-DC translation of Chinese DCs. Typically, English prepositions are tagged **'ALT'** and aligned to Chinese DCs that do not correspond with any English DCs. For example, ' 透過' is a common DC for the 'method' relation, yet there is not a DC for this relation in English and thus it is often translated to 'by' or 'through'.

4.4 Categorization of DCs

It is observed that subtly different DCs need not be distinguished for translation, thus they are annotated as variations of a same DC. For example, explicit occurrences of *'in addition'*, *'additionally'*, *'moreover'*, *'furthermore'* and *'besides'*, all listed as distinct DCs in PDTB, are annotated as instances of *'in addition'*, and ' 但是', ' 可是', ' 然而', ' 不過' as instances of ' 但是' (literally *'but'*). An unambiguous DC is used to represent the DC type, such as *'since'* as an instance of *'because'* but not the reverse.

Assigning DCs variations to an unambiguous type can serve as sense annotation without an abstract taxonomy of senses. External DC lexicon can also be flexibly added by registering new DC entries to existing categories. On the other hand, DCs that are not interchangeable in the syntactical context, such as *'but'* and *'however'*, are treated as distinct DC types in order to deduce discriminative translation rules.

4.5 Implicit DCs

In order to produce translation rules for all discourse relations, including the unmarked ones, implicit DCs (**IMP**) are inserted after all explicit DCs are identified in the Chinese EDU. A corresponding implicit DC is also inserted, if possible, as translation of a Chinese DC (explicit or implicit) when explicit translation is not identified. Note that implicit DCs are always annotated by a DC type instead of a variation to avoid ambiguity.

The **IMP** tag is used to annotate parallel DC structures in Chinese. Most Chinese discourse relations are marked by 'parallel DCs', which are similar to English patterns such as *'either...or'*, *'if...then'*, *'not only...but also'*. However, one or both DCs in the parallel structure can be dropped in Chinese. The dropped DCs are inserted as **IMP** and aligned to the English side.

After the first round of the annotation, another annotator is to repeat the annotation with the set of DCs recognized by the first annotator. Since implicit discourse relations lack lexical signals, the annotator agreement is lower (72% for English (Miltsakaki et al., 2004)). I plan to include implicit DC annotations of both annotators as multiple readings or coexisting DCs of the implicit relations, thus multiplying the training instances.

[1] A paragraph is considered an independent document in the PDTB. This annotation scheme follows this assumption.

4.6 Redundancy

Usually, two EDUs are related by one DC in English, thus only one of the Chinese parallel DCs is translated to explicitly. To learn this translation rule, the untranslated DC is thus aligned to a'**RED**undant' tag attached to the corresponding English EDU. To mark Chinese DCs that always occur independently rather than in parallel structure, the EDU without a DC is also annotated as '**RED**'. The various types of tags for DC annotation are summarized in Table 1.

Tags for aligned 'DC'

Chinese	English	
EXP	**EXP**	explicit DC identified
IMP	**IMP**	implicit DC insertable
-	**ALT**	expressions alternative to DC
RED	**RED**	ungrammatical to insert DC

Tags for Non-EDU Chinese segments

ATT	source of attribution
ADV	adverbial initialized
OPT	optional comma for a rhythmic pause

Table 1: Tags for Chi-Eng DC annotations

4.7 Primary analysis of the annotation

To date, 82 articles (about 33000 English words, about 1/3 of the complete dataset) have been annotated, giving rise to 2050 aligned discourse relations. In addition, 486 punctuation-separated segments on the Chinese side have been identified as non-EDU segments. 59 DC types for Chinese and 47 for English have been identified.

Chi -/- Eng	EXP.	ALT.	IMP.	RED.	Total
EXP.	291	68	23	49	**431**
IMP.	396	144	770	261	**1561**
RED.	6	0	0	52	**58**
Total	**693**	**212**	**783**	**362**	**2050**
attribute	-	-	-	-	211
optional	-	-	-	-	89
adverbial	-	-	-	-	186
Total	-	-	-	-	**486**

Table 2: Distribution of alignment between different 'DC' types

The distribution of alignments between these types is shown in Table 2. Although the statistics are not directly comparable to other existing data due to difference in definitions, it agrees with previous findings that implicit DCs are abundant in Chinese (Zhou and Xue, 2012). According to the present data, about 1/4 of the implicit DCs are translated to explicit DCs in English. However, more than half are not explicitly translated (implicit or redundant). This suggests that implicit DC recovery can be focused on the those that are likely to be translated explicitly.

It is also observable that explicit Chinese DCs are mostly translated to an explicit DC in English, while about 1/6 of them are translated to non-DC expressions. As mentioned, these are mostly prepositions corresponding to discourse relations that are not defined by any DCs in English. This suggests that bilingual discourse annotation can recover a larger variation of universal discourse relations than monolingual annotation. Further exploratory analysis will be conducted to investigate the tendency in discourse relation markedness and alignment, so as to define informative linguistic features for model training.

Currently, I am using the MAE annotation tool(Stubbs, 2011). The annotation effort can be lightened by developing an interface that assists the multilingual annotation task by, for example, automatic EDU segmentation (to be reviewed by annotators) and automatic identification and pre-alignment of DCs based on a DC dictionary.

5 Future plans

The key of this research is to integrate the annotated discourse knowledge into an SMT system. Integration of document level parse to MT, as described in Marcu et al. (2000) for Japanese-to-English translation, is complicated. In addition, comparing with Japanese, the word order in Chinese and English are not drastically different. Therefore, I plan to make use of information from DC-based shallow discourse parse. My main tasks towards this system include:

1. Cross-lingual DC annotation
2. EDU segmentation
3. Prediction of source implicit DCs
4. Integration to SMT system
5. DC-aware MT evaluation

A flowchart of these tasks is shown in Figure 1 and explained in the following.

5.1 EDU segmentation

Discourse parsing can be divided to the tasks of DC identification and argument identification,

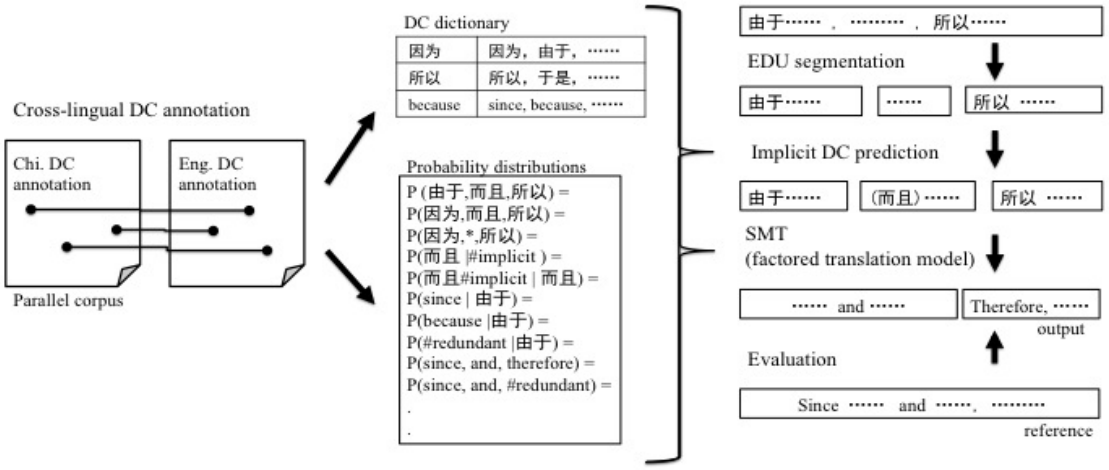

Figure 1: Main tasks for proposed DC-aware SMT system.

where the latter can be further divided into argument position and argument span identification. In Chinese, a punctuation-separated segment is basically considered an EDU, so the span is fixed. The exceptional cases of commas not segmenting an EDU are annotated in the dataset and can be predicted in a binary classification task using lexical and syntactical features, as in Yang and Xue (2012). On the other hand, a text segment can contain more than one EDU when there are multiple DCs, thus further segmentation is necessary depending on DC identification.

5.2 Prediction of source implicit DCs

One focus of this research is to explicitize implicit Chinese DCs when translating to English. I plan to construct a model to predict implicit discourse relations in the Chinese source text. Previous works on Chinese discourse relation recognitions (Yue, 2006; Huang and Chen, 2011) provide insights on the prediction task and the DC annotated corpus provides data for supervised training. Although state-of-the-art implicit discourse parsing is still of low accuracy, the preciseness can be adjusted to suit the goal of machine translation. As in other joint tasks with MT, such as Bouamor et al. (2013), features of whether the implicit DC can be translated explicitly, or correctly, can in incorporated to the prediction task, so as to predict translatable implicit DCs in particular.

5.3 Integration to SMT system

One way to exploit discourse knowledge into an SMT system is to incorporate the predicted discourse features, such as implicit DC, DC sequence or DC type, into a factored translation model (Koehn and Hoang, 2007). Another approach is to decorate identified and predicted DCs in a syntactical parsed tree, so as to enrich the tree-to-string rules with DC markedness features. Moreover, when a source DC is translated to a sentence initial DC, a source sentence is potentially split to multiple target sentences. A document level decoder (Hardmeier et al., 2012) that searches beyond the sentence boundary is thus preferred.

5.4 DC-aware MT evaluation

Comparable evaluation is essential for MT research, yet conventional MT metrics, such as BLEU, is not effective in detecting improvement in discourse relation translation (Meyer et al., 2012). One direction is to extend the *ACT* metrics (Hajlaoui and Popescu-Belis, 2013) to access also translation of implicit DCs. Another direction is to define a measure that is not reference-dependent, since implicit relations can be translated in various ways. Moreover, conventional MT metrics, which compare a candidate with the reference sentence-by-sentence, have to be modified when used to access the overall MT performance of the proposed system, since the output sentences may not align with the reference sentences one-by-one.

6 Conclusion

In this thesis proposal, ongoing work and future plans have been presented towards a discourse-relation-aware SMT system. The research can serve as basis for the goal of a document-level MT system that considers various discourse structures.

Acknowledgement

I would like to thank Baidu for travel and conference support for this paper.

References

Ann Bies, Martha Palmer, Justin Mott, and Colin Warner. 2007. English chinese translation treebank v 1.0. Linguistic Data Consortium LDC2007T02, January.

Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2013. Sumt: A framework of summarization and mt. *Proceedings of the International Conference on Natural Language Processing*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Proceedings of the SIGdial Workshop on Discourse and Dialogue*.

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. 2013. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue and Discourse*, 4(2).

Najeh Hajlaoui and Andrei Popescu-Belis. 2012. Translating english discourse connectives into arabic: a corpus-based analysis and an evaluation metric. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*.

Najeh Hajlaoui and Andrei Popescu-Belis. 2013. Assessing the accuracy of discourse connective translations: Validation of an automatic metric. *Computational LInguistics and Intelligent Text Processing*, 7617.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese discourse relation recognition. *Proceedings of the International Conference on Natural Language Processing*.

Hen-Hsen Huang and Hsin-Hsi Chen. 2012. Contingency and comparison relation labelling and structure prediction in chinese sentences. *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Hen-Hsen Huang, Chi-Hsin Yu, Tai-Wei Chang, Cong-Kai lin, and Hsin-Hsi Chen. 2013. Analyses of the association between discourse relation and sentiment polarity with a chinese human-annotated corpus. *Proceedings of the Linguistic Annotation Workshop and Interperability with Discourse*.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Ziheng Lin, Hwee Tou Ng, and Min Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. Technical report, National University of Singapore.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*.

Daniel Marcu, Lynn Carlson, and Maki Watanabe. 2000. The automatic translation of discourse structures. *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Thomas Meyer and Lucie Polakova. 2013. Machine translation with many manually labeled discourse connectives. *Proceedings of the Discourse in Machine Translation Workshop*.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Thomas Meyer, Andrei Popescu-Belis, and Najeh Hajlaoui. 2012. Machine translation of labeled discourse connectives. *Proceedings of the Biennial Conference of the Association for Machine Translation in the Americas*.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. *Proceedings of the Workshop on Frontiers in Corpus Annotations*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*.

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. 2012. Discourse-level annotation over europarl for machine translation: Connectives and pronouns. *Proceedings of the Language Resource and Evaluation Conference*.

Rashmi Prasad, Nikhit Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. *Proceedings of the Language Resource and Evaluation Conference*.

Ted Sanders and Leo Noordman. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 1.

Evgeny A. Stepanov and Giuseppe Riccardi. 2013. Comparative evaluation of argument extraction algorithms in discourse relation parsing. *Proceedings of the International Conference on Parsing Technologies*.

Amber Stubbs. 2011. Mae and mai: lightweight annotation and adjudication tools. *Proceedings of the Linguistic Annotation Workshop*.

Mei Tu, Yu Zhou, and Chengqing Zong. 2013. A novel translation framework based on rhetorical structure theory. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: a corpus-based analysis. *Computational Linguistics*.

Nianwen Xue. 2005. Annotating discourse connectives in the chinese treebank. *Proceedings of the Workshop on Frontiers in Corpus Annotations*.

Yaqin Yang and Nianwen Xue. 2012. Chinese comma disambiguation for discourse analysis. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Ming Yue. 2006. Discursive usage of six chinese punctuation marks. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*.

Yuping Zhou and Nianwen Xue. 2012. Pdtb-style discourse annotation of chinese text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. *Proceedings of the International Conference on Computational Linguistics*.

Lan Jun Zhou, Wei Gao, Binyang Li, Zhongyu Wei, and Kam-Fat Wong. 2012. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. *Proceedings of the International Conference on Computational Linguistics*.