Translating Implicit Discourse Connectives Based on Crosslingual Annotation and Alignment

Hongzheng Li[1], Philippe Langlais[2], Yaohong Jin[3]
[1] Institute of Chinese Information Processing, Beijing Normal University, Beijing, 100875, China
[2] RALI Lab, University of Montreal, Montreal, QC H3T 1N8, Canada
[3] Beijing Ultrapower Software Co., Ltd., Beijing, 100107, China
lihongzheng@mail.bnu.edu.cn, felipe@iro.umontreal.ca, jinyaohong@hotmail.com

Abstract

Implicit discourse connectives and relations are distributed more widely in Chinese texts, when translating into English, such connectives are usually translated explicitly. Towards Chinese-English MT, in this paper we describe cross-lingual annotation and alignment of discourse connectives in a parallel corpus, describing related surveys and findings. We then conduct some evaluation experiments to testify the translation of implicit connectives and whether representing implicit connectives explicitly in source language can improve the final translation performance significantly. Preliminary results show it has little improvement by just inserting explicit connectives for implicit relations.

1 Introduction

Discourse relations refer to various relations between elementary discourse units(EDUs) in discourse structures, these relations are usually expressed explicitly or implicitly by certain surface words known as discourse connectives(DCs).

Distribution of DCs varies between different languages. Let's just take Chinese and English for example. According to previous surveys, explicit and implicit DCs account for 22% and 76% respectively in the Chinese Discourse Treebank(CDTB) (Zhou and Xue, 2015), while they account for 45% and 40% in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), indicating that there are more implicit DCs in Chinese, correspondingly, discourse relations are usually implicit.

DCs should have some impacts on the translation performance and quality. As Chinese tends to use more implicit DCs, such DCs will be expressed explicitly when necessary in Chinese-English translation. Here is an example sentence show the implicit relation.

天气预报 说 今天 会 下雨，
weather report say today will rain
"Weather report says it will rain today,"
我们 决定 不 在 公园 举办 演唱会。
We decide not in park hold concert
"We decide not to hold the concert in the park."

There is no explicit DC between the two Chinese sub-sentences in the simple example, and the implicit discourse relation is CAUSAL. While translating into English, it is better to add an explicit DC such as "so/thus" before the second sub-sentence to express the relation, which will also make the translation more fluent and more acceptable.

In this paper, based on bilingual corpus, we first present cross-lingual annotation of DCs on both cross-sentence and within-sentence levels, and describe some related findings, then make a further survey on how to translate implicit DCs in Chinese-English discourse-level MT, and whether translation of DCs will have some impacts on final MT outputs.

The rest of the paper are organized as follows: section2 introduce some related works. Section 3 present annotation and findings of DCs in the bilingual parallel corpus. Section4 discuss some preliminary experiment results and analysis. And last section follow the conclusion.

2 Related Work

Discourse related issues have become increasingly popular in Natural Language Processing in recent years, especially the release of some famous discourse treebanks including PDTB, CDTB and RST (Mann and Thompson, 1986) corpus has promoted the research greatly.

Some research (Li et al. 2014, Rutherford and Xue, 2014) has done on monolingual annotation and analysis of Chinese DCs. Li et al. (2014a) and Yung et al. (2015a, 2015b) also present some cross-lingual discourse relation analysis. But they just analyze within sentences instead of cross-sentences.

In the field of MT, some previous works have been mainly focus on DCs in European language pairs (Becher, 2011; Zufferey and Cartoni, 2014) such as English, French and German, including but not limited to disambiguating DCs in translation (Meyer et al., 2011; Meyer and Popescu-Belis, 2012), labeled and implicit DCs translation (Meyer and Webber, 2013).

As for Chinese discourse relations and translation, Tu et al. (2013) employ a RST-based discourse parsing approaches in SMT, in their following work (Tu et al. 2014), they also present a tree-string model on Chinese complex sentences, integrating discourse relations into MT, gaining some improvement on translation performance. Li et al. (2014b) argues the influence of discourse factors in translation.

3 Cross-lingual Annotations of DCs

In order to investigate the DCs in the translation, we first manually align DCs in Chinese and English in the bilingual corpus, News-commentary corpus[1] downloaded from OPUS[2] (Tiedemann, 2012), then further annotate them with essential information on both the source and target sides.

The reasons why we choose news-commentary corpus lie in two sides: first, each line in the corpus usually includes several consecutive sentences, and each sentence is further composed of several sub-sentences(clauses), which provide rich cross-sentence and within-sentence discourse-level information. Second, sentences in each line are neither too long nor too short, which are suitable to train the MT models.

In this part, we will describe the annotation scheme and some corresponding findings.

3.1 Annotation Principles

As mentioned above, we will analyze the DCs on both cross-sentence and within-sentence levels, we decide to annotate the corpus in a top-down way. That is, we first annotate DCs between cross-sentences, and then within the sentences. Note that, if there exist sentences end with only full stop marks and have no commas or other punctuations, these sentences will not be annotated. Because they have no sub-sentences, and have no corresponding discourse relations within the sentence.

Here is an example:

a. 瑞典 本月 担任 欧盟 轮值 主席
Sweden this month as EU rotating presidency
有助于 推动 这项 计划。
help promote this plan

b. 但是 此时 正值 欧盟 东部 邻国
but now is EU eastern neighbor countries
面临 严重 挑战 的 时刻，因此 很多
face severe challenges DE time, so many
伙伴国 都 遭受 了 金融
partner countries all encounter financial
和 经济危机 的 沉重 打击。
and economic crisis DE severe hitting.

(Sweden's assumption of the EU Presidency this month should help these efforts. However, it comes at a time when the Union's eastern neighborhood faces severe challenges, and the financial and economic crisis hitting many of the partner countries hard.)

The example has two consecutive sentences *a* and *b*, we first need to indicate the DC and relation between them. Next, we will continue to analyze in *b*. As sentence *a* has no sub-sentences, we don't need to analyze on it.

Based on the principle, we first randomly extract 5,000 cross-sentence pairs from the corpus by using systematic sampling approach, and then extract possible sentences from the pairs.

Note that, as quite preliminary research, all current annotation is done by the first author of the paper alone, who is a PhD student majored in Linguistics and Computational Linguistics. As a result, unlike many previous works on corpus annotation, we don't conduct consistency experiments between different annotators to justify the performance of annotation until now. But we try to guarantee the annotation quality as much as possible. In the future, we will expand the annota-

[1] http://opus.lingfil.uu.se/News-Commentary.php
[2] http://opus.lingfil.uu.se/

tion size, asking other annotators to work together on the corpus and minimize the inconsistence during the annotation.

3.2 Annotation Labels

Inspired by (Yung, et al., 2015b), in our annotation scheme, we design several following labels. Most labels will be annotated on both cross/within-sentence levels on bilingual sides.

Nature of relations. Indicating the relations belong to explicit (E) or implicit (I) relations.

Explicit DCs. Annotating explicit DCs(EDCs) appeared in the sentences. On Chinese side, we try to find out all the possible DCs as much as possible. As for English, the DCs are annotated based on the 100 distinct types of explicit connectives in PDTB.

Implicit DCs(IDCs). If there are no explicit connectives in the sentences, proper DCs are inserted according to the discourse relations. If insertion is not grammatical, the DC is labelled as ‘redundant’.

AltLex. This label is only for English side, referring relations a discourse relation that cannot be isolated from context as an explicit DCs.

Semantic types of discourse relations. Considering the expression features of Chinese, based on the 8 senses of relations defined in CDTB, we also add 5 other relation types on Chinese side (shown in following table). As on English side, we adapt 4 top-level discourse senses defined in PDTB, namely Expansion(EXP), Contingency (CON), Comparison (COM) and Temporal(TEM).

Causation	*Purpose*
Conditional	*Temporal*
Conjunction	*Progression*
Contrast	*Expansion*
hypothetical	concession
example	explanation
successive	

Table 1: Relation types in Chinese. In which first 8 italic relations are defined in CDTB, and last 5 are newly added.

	Cross-sent (a, b)		**Within-sent (b)**	
	Zh	En	Zh	En
Nature	E	E	E	E
EDC	但是	however	但, 因此	However, and
IDC	/	/	/	/
types	Contrast	Com	Conjunction, Cause	Exp, Con

Table 2: An annotation example

According to the scheme, annotation of the above example in section 3.1 is shown in above table.

3.3 Annotation Statistics

Through the annotation, we annotate 5,000 cross-sentences and 8163 sentences, finally getting 5000 pairs of cross-sentence and 9308 within-sentence relations.

	Cross-sentence			**within-sentence**		
	Exp.	Imp.	Alt.	Exp.	Imp.	Alt.
ZH	1163 (23%)	3837 (77%)	/	2513 (27%)	6795 (73%)	/
EN	1094 (22%)	3622 (72%)	284 (6%)	4128 (44%)	4458 (48%)	742 (8%)

Table 3: Bilingual distribution of explicit and implicit relations

EN ZH	Exp.	Imp.	Alt.	Total
Exp.	947 (81%)	118 (10%)	88 (9%)	1163
Imp.	147 (4%)	3494 (91%)	196 (5%)	3837
Total	1094	3622	284	5000

Table 4: Cross-sentence DCs Alignment matrix

EN ZH	Exp.	Imp.	Alt.	Total
Exp.	1884 (75%)	351 (14%)	278 (11%)	2513
Imp.	2244 (33%)	4107 (60%)	464 (7%)	6795
Total	4128	4458	742	9308

Table 5: Within-sentence DCs Alignment matrix

Table3 shows on cross-sentence level, there exist more implicit DCs both in Chinese and English. The discourse relation “Consecutive” occupies highest frequency. While on within-sentence level there are still more implicit DCs than explicit ones in Chinese, but in English, their proportions are similar. The bilingual distribution of DCs in news-commentary corpus once again prove the similar findings in CDTB and PDTB before. We can also conclude that discourse relation types are more various within sentences, on the other hand, relations between sentences seem not so close, sentences are often independent with each other.

From the DC alignment matrixes in Table4 and 5, most explicit Chinese DCs usually have corresponding explicit DC translations. As for implicit DCs, although most of them map to implicit DCs on English side, there are still about 30% of them are aligned to explicit ones, indicating the important status and common usage of explicit DCs in English discourse structures.

We also find a quite prominent and interesting phenomenon that, a range of implicit discourse relations in Chinese, such as *Temporal, Conjunction, Coordination* and *Causation*, all can be mapped to the simple explicit DC “and” in English, with a rather high frequency. Just as similar conclusion shown in Appendix A of the PDTB 2.0 Annotation Manual[3], as one of top ten polysemous DCs, “and” can represent more than 15 senses in 3000 sentences in PDTB.

4 Preliminary Experiments & Analysis

We conduct MT automatic evaluation experiments on the annotated Chinese sentences with inserted implicit DCs to testify the translation performance before and after representing implicit DCs with explicit ones. Evaluation metrics include BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) scores, calculated by the Asiya toolkit[4] (Giménez and Màrquez, 2010).

4.1 Experimental Setting

With Moses decoder (Koehn et al., 2007), we train a phrase-based SMT model on another different version of News-Commentary corpus[5] provided respectively by OPUS (69,206 sentence pairs) and WMT2017 Shared Task[6] (235,724 pairs), and the model is tuned by MERT (Och, 2003) with the development sets (2002 pairs) provided by WMT2017. GIZA++ (Och and Ney, 2003) is used for automatic word alignment and a 5-gram language model is trained on English Gigaword (Parker et al., 2011). 1500 sentences randomly chosen from the annotated corpus in section3 are used as test sets.

The training data is not annotated with any discourse information, and thus the translation models are not trained with any discourse markups. But as the training data include both explicit and implicit DCs, it is suitable for the experiments.

4.2 Experimental Results and Analysis

	BLEU	METEOR
Before inserting implicit DCs	21.41	34.57
After inserting implicit DCs	21.43	34.56

Table 6: Evaluation scores of MT outputs

Table 6 shows the scores for SMT outputs of the test sets without/with inserting implicit DCs for source language. The scores indicate that adding explicit DCs for implicit DCs in Chinese seems have little improvement and impacts on translation performance.

We guess one reason resulting in the scores is that, although DCs appear frequently in English, they usually occupy a very small portion of total word counts in the MT outputs and may not very sensitive to BLEU. As (Meyer et al., 2012) also argues that translation of DCs can actually be improved while BLEU scores remain similar.

After manually analyzing some sentences of the outputs, it is observed that after inserting explicit DCs for implicit relations, most of them are indeed translated and aligned to the source side, just as the examples shown in following table7, stating that our preprocessing for the implicit DCs can be identified by the decoder. But, if we compare the translated DCs with those in reference, some of them are different, thus the n-gram based BLEU evaluation will not able to capture the information, which support our guess.

Source: 作为货币联盟，金融一体化在欧元区非常牢固， [implicit = Causation, added DC = ***因此***] 这使得欧洲央行成了不二之选。 **Ref:** Given that financial integration is particularly strong within the monetary union, putting the ECB in charge was an obvious choice. **MT:** As a monetary union, financial integration in the euro area is very strong, ***so*** it makes the ECB has become the best choice.
Source: 这些国家需要采取措施助贫民摆脱贫困陷阱， [Implicit = Coordination, added DC = ***并且***] 给他们现实的机会改善其经济福利。 **Ref:** These economies need measures that help to keep the poor out of poverty traps, and that give them realistic opportunities to improve their economic well-being.

[3] https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf
[4] http://nlp.lsi.upc.edu/asiya/
[5] http://opus.lingfil.uu.se/News-Commentary11.php
[6] http://www.statmt.org/wmt17/translation-task.html

MT: These countries need to take measures to help the poor get rid of poverty traps ***and*** give them real opportunities to improve their economic well-being.

Table 7: Some examples of MT outputs

5 Conclusion

In this paper, we cross-lingually annotate and align DCs from both the cross-sentence and within-sentence levels on a Chinese-English parallel corpus. Based on the annotation, we present some statistics and basic findings on DCs, which have some accordance with previous survey.

We also conduct some preliminary MT evaluation experiments to testify the impacts on translation performance resulted from expressing implicit DCs explicitly. Although the results temporarily indicate no significant improvement of MT outputs, preprocessing DCs for MT indeed has some positive effects, we still believe that DCs are one of useful factors that cannot be ignored for discourse-level MT.

In the future, we need to consider other possible discourse-related information and integrate them into MT, on the other hand, it is also worthy considering more on the issue that how to evaluate discourse-MT outputs properly, after all, BLEU scores alone may not enough.

Acknowledgments

We would like to sincerely thank all the reviewers for their constructive and helpful suggestions on this paper. The work is supported by China Scholarship Council (Grant No. 201606040144).

References

Viktor Becher. 2011. When and why do translators add connectives? a corpus-based study. *Target*, 23(1): 26 –47.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In: *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*, pages228-231.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014a. Cross-lingual discourse relation analysis: A corpus study and a semi-supervised classification system. In: *Proceedings of the International Conference on Computational Linguistics*, pages577-587.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014b. Assessing the discourse factors that influence the quality of machine translation. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages283-288.

Yancui Li, Wenhi Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014. Building Chinese discourse corpus with connective-driven dependency tree structure. In: *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages2105–2114.

William C Mann and Sandra A Thompson. 1986. Rhetorical structure theory: Description and construction of text structures. *Natural language generation: New results in artificial intelligence, psychology, and linguistics*, 279-300.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey et al. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In: *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue,* pages194-203.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In: *Proceedings of the Workshop on Hybrid Approaches to Machine Translation*, pages 129-138.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In: *Proceedings of the Discourse in Machine Translation Workshop.*

Jörg Tiedemann, 2012. Parallel Data, Tools and Interfaces in OPUS. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214-2218.

Philipp Koehn, Hieu Hoang, Alexandra Birch et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages177-180.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19-51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics,* pages 160-167.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword fifth edition ldc2011t07. *Linguistic Data Consortium*.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of 40th Annual Meeting of the ACL*, pages 311–318.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In: *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages645–654.

Mei Tu, Yu Zhou, and Chengqing Zong. 2013. A novel translation framework based on rhetorical structure theory. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 370-374.

Mei Tu, Yu Zhou, and Chengqing Zong. 2014. Enhancing grammatical cohesion: Generating transitional expressions for smt. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 850-860.

Frances Yung, Kevin Duh, and Yuji Matsumoto. 2015a. Sequential annotation and chunking of Chinese discourse structure. In: *Proceedings of The SIGHAN Workshop on Chinese Language Processing*, pages1-6.

Frances Yung, Kevin Duh, and Yuji Matsumoto. 2015b. Crosslingual Annotation and Analysis of Implicit Discourse Connectives for Machine Translation. In: *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT)*, pages 142–152.

Sandrine Zufferey and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target*, 26(3):361 –384.

Yuping Zhou and Nianwen Xue. 2015. The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397-431.