Towards the Design of Cross-lingual Discourse Annotation in Chinese-English Bitext

Frances Yung　Kevin Duh　Yuji Matsumoto
Nara Institute of Science and Technology
{pikyufrances-y,kevinduh,matsu}@is.naist.jp

1 Introduction

State-of-the-art MT systems translate individual sentences in an article independently, and research on MT tasks beyond sentence level are considered as high-hanging fruits. Nonetheless, human translation, being the training data for MT as well, is often 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 chain, and lexical coherence.

Among various discourse structures, we focus on the transfer of discourse relations, which are the connections between units of concepts in a text. These connections turn a set of individual sentences to a coherent discourse with an overall meaning. Discourse relation markers in a text sometimes map ambiguously with the underlying discourse relations, while some relations are even unmarked. A reasonable initial attempt to learn discourse-relation-awared translation rules is to explicitly learn from the underlying discourse relations of a translation corpus. Towards this aim, we propose to design a scheme that annotates marked and unmarked discourse relations in a Chinese-English translation corpus and aligns them cross-lingually.

Section 2 gives an overview of existing literature on discourse for MT. Section 3 describes our annotation scheme in detail. Finally, a conclusion is drown in Section 4.

2 Related work

Research in discourse processing typically base on discourse-annotated corpora, among which the Penn Discourse Treebank (PDTB) [9] is of the largest scale. Its lexically ground annotation associates underlying discourse relations with surface words known as *'discourse connectives'* (DC), which can either be actually present (i.e.*explicit*) in the text or absent (i.e.*implicit*). An example is shown below.

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 also annotated with predefined *senses* classified into 3 levels of granularity.

PDTB allows evaluation of English discourse parsing tasks and disambiguation tasks [7, 2], which reveal that implicit discourse relations are much harder to learn compared to explicit discourse relations [6, 16]. On the other hand, schemes for Chinese discourse annotation have been proposed in the existing literature [12, 15] but the corresponding resource is not yet available.

Earlier studies of discourse relations in MT includes [4], which proposed a discourse transfer model to re-construct the target discourse tree from the source discourse tree, parsed by the Rhetorical Structure Theory [3] (RST), a traditional discourse 'grammar'. 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 with its pre-defined sense 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 [5, 8]. Experiments of these works have been conducted in English-to-French, Czech and German translation and only explicit DCs were considered. Tu et al. [11] 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, but only intra-sentential, explicit relations are considered.

3 Cross-lingual annotation of discourse relations

Motivated by related works of learning discourse relations from annotation, we propose to annotate and align DCs in a Chinese-English corpus. In particular, a much larger proportion of discourse relations is implicit in Chinese [12], and thus many implicit relations have to be explicitly translated in English. Moreover, since discourse relations relate *elementary discourse units* (EDUs) inside and outside a sentence, discourse-relation-awared translation model allows EDU-to-EDU translation instead of sentence-to-sentence translation, which intuitively suits the translation between Chinese and English. It is because a Chinese sentence typically consists of a list of semantically-related fragments, while an English sentence is a strict, parsable syntactic unit. A model that translates at EDU level and joins the EDUs by discourse relations also helps joining fragments by connectives or splitting long sentences to shorter ones. Towards such a model, we design a formalism to align the underlying discourse relations in a bilingual corpus.

The data we use is the English Chinese Translation Treebank [1], which consists of 325 Chinese news stories translated into 146,300 words of English. The annotation of discourse relation is based on the lexical approach of PDTB, with each relation signaled by a DC. The transfer of discourse relation is annotated by aligning each DC from Chinese to English, while arguments and senses are not annotated. Some annotation examples are shown below. The design is based on the PDTB-styled annotation and translation spotting of DCs, with adaptation to capture the cross-lingual characteristics of discourse structures, which are discussed in later subsections.

Example 2 新建的骨科、婦科、兒科三個專科醫院 , (**optional**) 設備先進, (**align 1**, *implicit DC*=和) 已開門應診。
The three recently constructed hospitals specializing in orthopedics, gynecology and pediatrics have advanced equipment and (**align 1**, *explicit DC=and*) are open to patients. (CTB 020)

Example 3 據介紹 , (**attribution**) 這十四個城市的城市建設 和 (**not annotated**) 合作區開發建設步伐加快。(**align 1**, *implicit DC* =具體來說) 三年來 , (**adverbial**) 這些城市累計完成固定資產投資一百二十億元, ...
According to a briefing, the pace of municipal construction and(**not annotated**) of construction for opening of the cooperation zones of these fourteen municipalities has accelerated. (**align 1**, *implicit DC=specifically*) In the past three years these municipalities collectively have put together investment in fixed assets in the amount of 12 billion yuan. (CTB 003)

Example 4 (**align 1**,*implicit DC*=其實) (**align 2**,*implicit DC*=雖然) 在投資項目上比上年減少四百四十四件, 但 (**align 3**, *explicit DC*=但是) 投資金額 卻 (**align 4**, *explicit DC*=卻) 比上年增加一點三億多美元。
(**align 1**, *implicit DC=in fact*)(**align 2**, *redundant*) The number of investment projects dropped by 444 as(**not annotated**) compared with last year, but(**align 3**, *explicit DC=but*) the value of investments (**align 4**, *redundant*) rose by more than 130 million US dollars as(**not annotated**) compared with last year. (CTB 012)

Example 5 在進行全球貿易自由化的同時 (**align 1**, *explicit DC*=同時), (**align 2**, *redundant*) 中國必須對國有企業進行改革, (**align 3**, *implicit*=以) 增強本身的競爭力。
While(**align 1**, *explicit DC=while*) implementing global trade liberalization, (**align 2**, *redundant*) China must implement reforms on state-owned enterprises so as to (**align 3**, *explicit DC=in order to*) improve its own competitiveness. (CTB016)

3.1 EDU segmented by punctuations

Chinese sentences are usually long and separated into segments by punctuations. From the viewpoint of discourse structure, each comma-separated segment can be considered as an EDU [13, 15] and can be aligned across the two languages. The punctuations separating the EDUs are strong clues for the identification of discourse relations[14]. Nonetheless, not all segments separated by commas are EDUs, since Chinese commas are used arbitrarily to signify '*pauses*' in the sentence. Exceptions to be tagged to the commas not acting as EDU segmenters include markers of 'attribution', 'initial adverbial', and optional commas placed after a long subject (examples 2, 3). The annotation can be used to train automatic classifier of EDU segmenting punctuations[13].

3.2 Explicit DCs

A list of 100 DCs are defined in PDTB, but they are annotated only if they are used to relate '*abstract objects*', which are typically clauses. For example, the '*and*' in Example 2 is not considered as a DC, but would be annotated as one if *Arg2* is a clause (e.g. *'and they are open to patients.'*). If the same restriction is applied to Chinese, a large number of discourse relations will be excluded from annotation since subjects are often dropped. Therefore, we propose to base on semantics rather than syntactic structure in DC identification and consider both cases of 'and' as DCs. Nonetheless, when the token is used to join two distinct entities or actions,

such as the '和' and '*and*' in Example 3, or when two arguments are not identifiable, such as the '*as*' in Example 4, it is not annotated as DC nor aligned.

DCs defined in PDTB are either *subordinating conjunctions*, *coordinating conjunctions*, or *adverbials*. In cross-lingual annotation, however, a typical DC in one language may not be translated as a target DC by the strict definition. In order to improve the coverage of cross-lingual annotation and capture more translation rules, we propose not to restrict on the syntactic category as long as a word or a multiword expression functions as a DC. For example, 'on the other hand', 'at the same time', and 'in spite of' are all annotated as DC instances, while in the PDTB, they are annotated as DC, annotated only as implicit DC, and not annotated respectively.

3.3 Categorization of DCs

We notice that some DCs annotated in PDTB differ very subtly. For example, *'in addition'*, *'additionally'*, *'moreover'*, *'furthermore'* and *'besides'* are listed as distinct DCs in PDTB, but basically any of them can be used to mark the same discourse relation. Similarly, the Chinese DCs '可是', '但是', '然而', '不過' can all be grossed to *'but'*. Therefore, similar DCs are annotated as instances of the same DC type in our scheme, since it is not necessary to distinguish interchangeable source DCs during translation and grouping them into categories helps reduce data sparseness. For example, instances of the above 4 Chinese DCs are all annotated as variations of '但是', a frequent and unambiguous DC that marks the contrast relation[1]. In particular, Chinese words can often be abbreviated, such as '但' for '但是' in Example 3. Such abbreviation can be the cause of ambiguity. For example, 而 is ambiguous since it can be the abbreviated form of '而且' (and), '因而' (therefore), '然而' (but), and '反而' (instead). Mapping the variants to the DC category has the same effect as annotating them with senses, yet an abstract texonomy of senses need not be pre-defined. External DC lexicon can also be flexibly added by registering new DC entries to existing categories.

Nonetheless, the categorization of DCs is restricted by their syntactical usage and members of a DC type has to be interchangeable semantically and syntactically. For instance, *'but'* and *'however'* belong to 2 distinct DC types since *'but'* cannot be used in the beginning of a sentence as *'however'* can. Similarly, '卻' cannot be inserted in front of the subject as '但是' can (Example 4). Such distinction is useful for extracting the arguments of the DC, as some DC require taking an *Arg1* inside the same sentences while some in the previous sentences [10]. However, interchangeable structural variants (such as '的同時' for '同時' in Example 4) and DCs having more than one syntactic usages (such as sentence-initial and sentence-middle *'thus'*) are grouped under the same category. The interchangeability of the DCs is to be tested by accessing the acceptability of the sentence when the DC instance is replaced by the potential DC type it belongs, without considering stylistic or rhythmic differences. Note that the DC variants defined in our scheme is different from their definitions of *modified connectives* in PDTB, which are connectives modified by adverbs, such as *'partly because'* for *'because'*. Our method is similar to the approach taken by [15], in which implicit relations are annotated directly with senses associate with one or two prototypical DCs. However, our categorization is not grounded on any sense definition and is applied to explicit relations as well. Also, we assume that any discourse can be parsed into an RST tree, thus, in contrast with PDTB-styled annotation, DC-associated discourse relation always exists between two EDUs.

3.4 Implicit DCs and Chinese parallel structure

In English, there is a small set of 'parallel DCs', such as *'either...or', 'if...then', 'not only...but also'*. These are annotated in PDTB by defining the first half of the parallel structure as Arg1, and the second half as Arg2. Parallel structure, in contrast, is abundant in Chinese discourse. While some DCs always occur parallel, such as '一... 就' (*'once...then'*), other complete structures are often 'abbreviated' by dropping one of the DC pair. For example, the '雖然' of the'雖然... 但是' (*'although...but'*) pattern in Example 4 is dropped. In turn, an extra DC '卻' (yet) is used to mark the contrast relation. In fact, using either any one, two, or all three of these DCs signals the same discourse relation in Chinese and the variance arbitrarily depends on the tone or rhythm of the sentence. In English, however, it is redundant to use more than one of the three.

Zhou and Xue [15] suggests to annotate the arguments with senses as well, such as 'cause' and 'reason', instead of depending on the parallel or redundant DCs. Instead, we carry through the lexically grounded approach by deliberately inserting implicit DCs to associate each argument with a DC. For example, the *Arg1*-but-*Arg2* pattern is annotated as 雖然-*Arg1*-但是-*Arg2* in Chinese, where '雖然' and '但是' can either be implicit or explicit. Inserting an implicit 雖然 to *Arg1* has the same effect of argument annotation, given that EDUs are segmented by punctuation in Chinese.

Nonetheless, some DCs only allow independent usage rather than parallel structure. The *'redundant'* tag is annotated to the EDU when it is ungrammat-

[1] Although one may argue that *'besides'* alone is used in informal context, or that '然而' has a softer tone than others, we take formality and tones as semantic aspects independent of discourse relation.

ical to insert any DC (Example 5). Due to the difference in frequency, Chinese parallel DCs are often aligned to single DCs in English. In such cases, the extra DCs can be aligned to the *'redundant'* tag on the English side (Example 4).

4 Conclusion

Towards the goal of developing an SMT system which considers the transfer of discourse relation across languages, we propose to enrich a parallel corpus with cross-lingual discourse annotation. This paper discusses the design issues that are important to such an annotation effort. Lexically grounded representation of discourse relations can be more easily incorporated to an SMT system, hence we adopt an annotation scheme that avoids verbal definitions of DC or argument senses, but associates the senses to categorized implicit or explicit DCs. In this way, the cross-lingual gap in discourse structure lexicalization can be represented in terms of the sequence and explicitness of the underlying discourse relations. Our annotated corpus is currently under construction. Our final goal is to learn translation rules from the annotated data and develop a discourse transfer model that can be incorporated to an SMT system.

References

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

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

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

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

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

[6] Emily Pitler, Annie Louis, and Ani Nenkova. 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*, 2009.

[7] Emily Pitler and Ani Nenkova. 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*, 2009.

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

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

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

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

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

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

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

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

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