A Bilingual Discourse Corpus and Its Applications

Yang Liu[1], Jiajun Zhang[1], Chengqing Zong[1], Yating Yang[2] and Xi Zhou[2]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2]Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi, Xinjiang, China

{yang.liu2013, jjzhang, cqzong}@nlpr.ia.ac.cn

{yangyt,zhouxi}@ms.xjb.ac.cn

Abstract

Existing discourse research only focuses on the monolingual languages and the inconsistency between languages limits the power of the discourse theory in multilingual applications such as machine translation. To address this issue, we design and build a bilingual discource corpus in which we are currently defining and annotating the bilingual elementary discourse units (BEDUs). The BEDUs are then organized into hierarchical structures. Using this discourse style, we have annotated nearly 20K LDC sentences. Finally, we design a bilingual discourse based method for machine translation evaluation and show the effectiveness of our bilingual discourse annotations.

Keywords: discourse analysis, bilingual discourse parser, SMT metric with discourse information

1. Introduction

Discourse theory has been studied for decades. The task of discourse analysis is to segment sentences into non-overlap elementary discourse units (EDU) and then reorganize them via discourse relations to form discourse structures such as linear chains(Eisenstein and Barzilay, 2008), trees (Feng and Hirst, 2013) and graphs (Wolf and Gibson, 2005) in various discourse banks e.g. RST and PDTB. Due to the semantic integrity of EDUs, EDU relations and their well-formed structures, discourse knowledge has been applied to many natural language processing (NLP) tasks, such as information extraction, summarization, QA and statistical machine translation (SMT). Previous research works have proven that discourse information is beneficial to these NLP tasks.

However, the current discourse annotations concentrating only on monolingual languages have some insufficiency that limits its power in many multilingual NLP tasks. Let us take statistical machine translation for example. Machine translation aims at finding for the source language sentence a target language sentence, which shares the same meaning as its source side. Intuitively, the source sentence and its target translation should have the similar discourse structure. Several recent research works apply the monolingual discourse knowledge to SMT, source side or target side. Some of them enforce the target language translation to be consistent with the source-side discourse structure (Tu et al., 2014). The others attempt to measure the translation discourse structure using discourse structures of target language references (Joty et al.,2014). Since there are wide variations in discourse annotations between source language and target language, the above two methods are somewhat conflicting. The problem is similar in other multilingual NLP tasks e.g. multilingual summarization (Anechitei and Ignat, 2013) ,discourse translation (Marcu et al., 2000; Tu et al., 2013) and bilingual semantic role labeling (Yang et al., 2015).

Through the above analysis, it is obvious to realize that developing multilingual discourse corpus is quite necessary for multilingual NLP tasks and helpful to understand the diversity of discourse information in multilingual environment. Therefore, at the start point, we propose and annotate a bilingual discourse corpus. Furthermore, we design a MT evaluation metric based on the bilingual discourse annotations.

2. Discourse Inconsistency between Languages

First, to have a better understanding, we manually analyze the discourse structure difference between two languages. We use Chinese-to-English SMT evaluation test set NIST2003 as the dataset that is originally designed for SMT Task, including 919 Chinese sentences, and for each Chinese sentence there are 4 English reference sentences. We select the widely-used Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) to represent the discourse structure of the Chinese and English sentences respectively.

By comparing the results, we first observe that the four English references with same meaning usually share similar EDU segmentation, structure and relations, (the consistency is acceptable) as shown in Table 1.

Consistency	Segment	Structure	Relation
REF(0-1)	87.23%	58.91%	51.21%
REF(0-2)	86.12%	62.24%	51.94%
REF(0-3)	89.34%	60.67%	53.64%
REF(1-2)	83.68%	54.35%	41.11%
REF(1-3)	87.86%	57.61%	47.33%
REF(2-3)	88.89%	56.56%	43.18%

Table 1: The four English references generated in RST-style, and most of them can match with each other.

As we show in Figure 1, there are two sentences with exactly same meaning. Despite they have different discourse trees, they still share similar structure and

discourse relation. The difference between discourse trees of example sentences is caused by segmentation. For this example, they share a common EDU and a common discourse sub-tree. Usually, this kind of distinctions is insignificant.

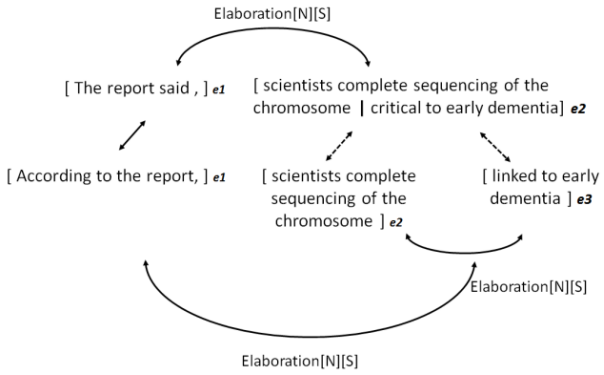

Figure1: Different discourse trees between references. These two sentences have the same meaning. Different segmentations lead to different discourse structures. Note that '|' implies the potential segment position.

However, when comparing the Chinese sentence with its English references, the situation is quite different. We find that there is a significant divergence of segmentation, discourse structure, and relations.

As we show in the Figure 2, the situation seems very similar to the example above, but actually, these two examples are different in essence. The Chinese chunk, which is aligned to the English reference edu3, is nested in another Chinese chunk ***edu2***.

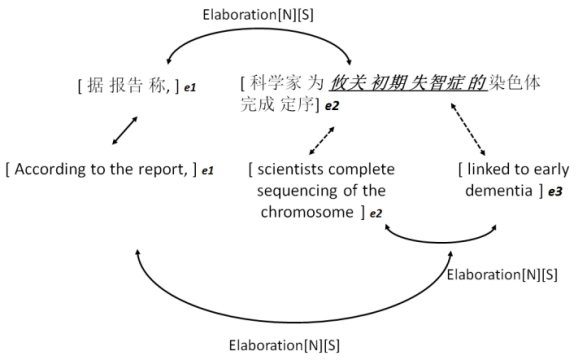

Figure 2: the nested structure in Chinese sentence. Unlike the previous example, there is no potential position that can split the Chinese ***edu2*** into two separate parts.

The segmentation score is computed with a simple way on Chinese and English respectively. For instance, the example in Figure 2 contains two Chinese EDUs and three English EDUs. They share a common EDU which is fully aligned, and then the segment score for Chinese is 1/2, and the score for English is 1/3. And only when segmentation is matched in both languages, we can calculate the match score of structure and relations (for convenience, we use the same discourse relation on Chinese sentence as English). This can give us direct perspective of discourse difference in these two languages. Table 2 shows the results. For every Chinese sentence, we select the most similar reference English sentence

Consistency	Segment	Structure	Relations
Test sentence	52.21%	25.39%	19.67%
Most similar reference	44.82%	20.81%	16.34%

Table 2: Comparison between the Chinese discourse parsing results and their closest English reference discourse parsing results.

We have done a deep investigation and we believe there are three potential issues that cause the big divergence of discourse structures:

- EDU Definition and Segmentation: in discourse framework, clause is usually considered as EDU, but the way to identify clause follows different principle in different languages even if they share similar semantic meaning.
- Structure: discourse structure links EDUs together, mainly based on the structure of a sentence. Therefore, discourse structure can be affected by inconsistency of expression style in each language.
- Relations: it is easy to determine the relationship between EDUs if there is a conjunction, which is also related to explicit relation. But the classification of implicit discourse relations still remains a difficult problem in discourse study (Li et al.,2015).

Since both of the structure and relations are based on the EDU segmentation, we mainly focus on definition and segmentation of bilingual elementary discourse units (BEDU) in our first version.

3. Bilingual Elementary Discourse Unit

Intuitively, the source and target side of a bilingual elementary discourse unit should be monolingual EDUs sharing the same meaning. From the opposite direction, there are three possibilities:

There exist a source language EDU and a target language EDU expressing the same semantics, and then they can form BEDU naturally.

There is a source language EDU, but its target part is not recognized as an EDU.

There is a target language EDU, but its source part is not identified as an EDU.

We have conducted a detailed analysis and try to figure out which case is more popular. According to our observation, we find that consistency of the EDU segmentation is very low under bilingual circumstance. The last two cases overwhelm the standard bilingual discourse structures. Figure 3 shows an example about a pair of parallel sentences, annotated with word alignment and part-of-speech. The span " linked to early dementia " is identified as EDU in English, while the span " 攸关 初期 失智症 " aligned to it is not considered as an EDU because this span is just embedded in another EDU.

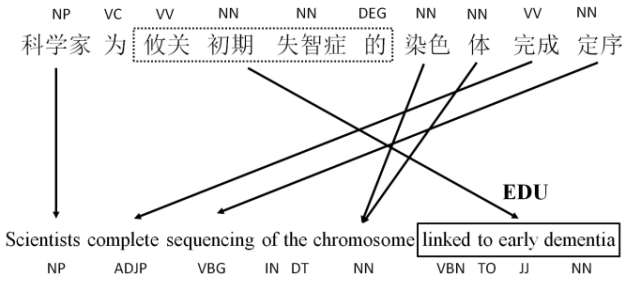

Figure 3: word alignment of parallel sentence, the English chunk in block is aligned to the Chinese chunk in dashed box

The case we show above occurs frequently under bilingual situation. When a span is recognized as an EDU in English, the Chinese span aligned to it may not be considered as an EDU and vice versa, even though they have the same meaning, grammatical function, and semantic function.

As we know that an EDU in a sentence has semantic integrity and is usually an exact substructure of syntactic tree. Therefore, to define bilingual elementary units, we aim at looking for a bilingual structure, which has similar function to monolingual EDUs and remains semantic, consistent in bilingual sentences.

After analyzing the parallel sentences, we notice that:

Semantic relation between words remains consistency in parallel sentence as long as they share the same meaning.

A verb can maintain a relatively stable substructure working as integrity in both languages.

This inspires us to make full use of dependency parse structures of parallel sentences with EDU segmentation generated by source and target monolingual discourse parsers. We find some interesting facts, and show them in Figure 4.

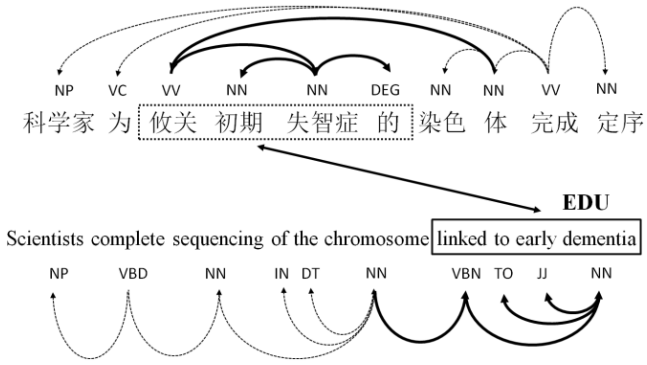

Figure 4: Structure of dependency and discourse segmentation. Apparently, these two chunks are matched in dependency sub-tree level.

As shown in Figure 4 that the monolingual EDUs match the dependency substructures dominated by a verb "linked" in English, and that dominated by verb "攸关" in Chinese. In addition to this example, we also find that this property exists in most of bilingual dependency parse tree structures.

The study indicates that in many circumstances monolingual EDUs always match the dependency substructure if the dominated word is a verb. This is to say a well-formed dependency substructure headed by a verb is usually semantic integrity. Thus, we define two types of BEDU as follows:

The first type is hard-BEDU, a source-side span and a target-side span compose a hard-BEDU if and only if the source and target span are translations with each other, are both well-formed dependency substructures (dominated by a verb1) and the source or the target span should be an EDU in the monolingual language.

For example, the source span "攸关 初期 失智症 的"and the target span "linked to early dementia" is a hard-BEDU.

Another type is soft-BEDU, as we see the example shown in Figure1, two English sentences have different discourse structures only because of the segmentation. We also define the soft-BEDU for this circumstance, which can improvement the consistency score in English sentence pair. The motivation behind is that the target side sentence variant can affect the consistency on bilingual circumstance. By improving the consistency score in target side, the source side can also benefit from it.

In brief, our philosophy of BEDU definition and segmentation is to choose the most stable structure which keeps semantic consistency in both languages and can be treated as an EDU in at least one language.

It should be noted that we can use a coarse-to-fine method to annotate the BEDUs. First, we develop an unsupervised approach using language model and word alignment to recognize the cohesive bilingual units. Then, we manually refine these bilingual units to form BEDUs. It saves us a lot of time for annotation.

4. Bilingual Discourse Structure

Once the bilingual elementary discourse units have been determined, the next step is to build the discourse hierarchical structure. Different from monolingual discourse trees, we have to build bilingual discourse structure (BDSs), one is for source language, and the other is for target language.

As mentioned above, the way we identify the BEDU actually indicates that we consider BEDU as a basic semantic component of sentences. Therefore, the BDSs reflect the structure of bilingual semantic components. Given the parallel sentences with BEDU segmentation shown below, the corresponding BDS is illustrated in Figure 5.

S: [这 是 **[**到 目前 为止 完成 定序 的**]** $_2$第四 对 染色体， **[**它 由 八千七百多万 对 去氧核糖核酸 (DNA) 组成 。 **]**$_3$ **]**$_1$

T: **[**this is the fourth chromosome **[**to be fully sequenced up till now**]**$_2$, **[**it comprises more than 87 million pairs of dna . **]** $_3$**]**$_1$

[1] We can loose this constraint and allow more BEDUs.

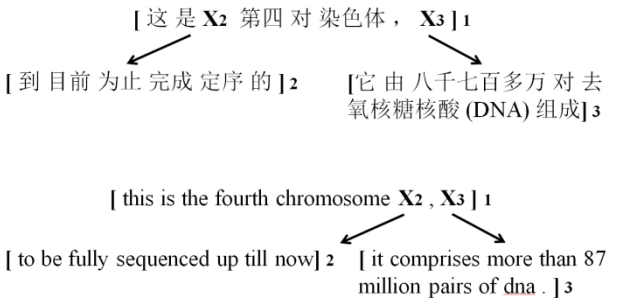

Figure5. The bilingual discourse structure building

We can see from Figure 5 that the bilingual discourse structure is different from the standard monolingual ones. In monolingual discourse trees, all EDUs are linear and there is no nested substructure. In our bilingual discourse structure, nested substructure is very normal just as shown in the Chinese part of Figure 5.

Also, for English sentence pair, we build BDS following similar process, and the only difference is that we introduce the soft-BEDU into structure. We show the BDS for English sentence pair in Figure 6.

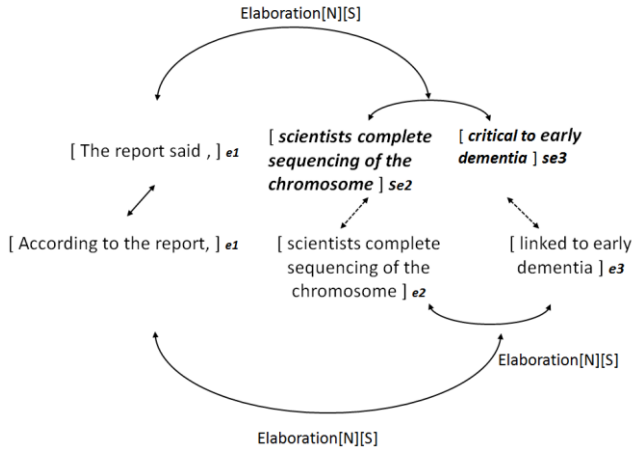

Figure 6: Same example in Figure1 but as we introduce the soft-BEDU the two sentence is matched now.

Those characteristic above poses a tough problem: how to annotate the discourse relations between BEDUs. From the definition of BEDU, we know that the nested substructure appears only on one language side. Therefore, we resort to the monolingual EDU relation to determine the BEDU relation. For example, the relationship of (X2_chen, X3_chen) is the same as that of (X2_en, X3_en). One important property of our BDSs is that we allow the existence of discontinuous BEDUs in the structure. Another is that we also introduce the soft-BEDU to improve English side consistency. Since discourse annotation is a time consuming job, we remain this task as our future work.

5. Preliminary Bilingual Discourse Corpus

In machine translation community, there are large-scale parallel sentence pairs. Thus, we first annotate bilingual discourse structures on this data. Considering the direct application of this BDT corpus, we choose Chinese-English SMT test sets (NIST2003, NIST2004 and NIST2005) for annotation. There are respectively 919, 1788 and 1082 Chinese sentences in NIST2003, NIST2004 and NIST2005. Each Chinese sentence has four English sentence references. So, we have 18,945 (919*5+1788*5+1082*5) parallel sentence pairs for annotation.

Using this annotated BDT corpus, we have conducted a detailed analysis to see whether the semantic components of parallel sentences are consistent for BEDU recognition. Table 3 gives the statistics. It shows that the consistency is very high between Chinese and English. Furthermore, this kind of bilingually-constrained BEDUs are more interpretable than the monolingual EDUs.

Consistency	Segment	Structure
REF(0-1)	98.23%	87.45%
REF(0-2)	97.86%	86.14%
REF(0-3)	98.12%	87.12%
REF(1-2)	97.95%	86.39%
REF(1-3)	98.18%	87.14%
REF(2-3)	98.03%	87.23%
TEST-REF0	94.20%	85.39%
TEST-REF1	93.81%	84.80%
TEST-REF2	94.04%	85.27%
TEST-REF3	93.90%	85.01%

Table 3: Consistency of Segmentation and Structure between Chinese and English sentence pairs.

We intend to apply our BDT ideas and the annotated BDT dataset in many NLP tasks, such as SMT (translation and evaluation), dependency parsing, multilingual summarization and cross-lingual information extraction.

Next, we introduce a direct application which uses the BDT corpus to design a discourse-driven evaluation metric for SMT.

6. SMT evaluation metric with Discourse information

We develop a simple evaluation method to show how to use our annotated corpus.

BLEU (Papineni et al.,2002) is the most popular SMT evaluation metric by now. It is a simple method based on string matching. However, this method is criticized for ignoring the linguistic characteristic of translation such as discourse information.

Since we have annotated the test set of SMT task, we can evaluate our translation output with discourse information. This novel discourse-driven SMT evaluation metric measures three key points: BEDU integrity, BDS structure, and BEDU coherence. We conduct an

experiment on our SMT task and the results show that our discourse-driven evaluation can benefit from discourse annotated corpus and has higher correlation score with human judgment.

Our metric hypothesis is that the similarity between discourse structures of an automatic and a reference translation provides additional information that can be valuable for evaluating MT systems. We now need the word alignment of each BEDU pair between source sentence and target sentence. The word alignment is learned from bilingual parallel corpus from LDC[2]. The word alignment tool is GIZA++ (Och, 2000) and the word alignment are symmetrized using the grow-diag-final-and heuristic.

Then we can measure BEDU integrity score by measuring the aligned rate for each BEDU pair from source side and target side. In addition, we can measure the translation sentence discourse structure score by aligned rate from reference EDU to translation output. We also collect conjunction word set from PDTB to measure the discourse coherence score. We combine those three score metric features with original BLEU metric as our discourse sensitive metric. The word alignment can be learned from bilingual parallel corpus, which is easy to come by.

An ideal SMT evaluation metric should be able to rank the translation results from bad to good and achieve high correlation with human judgment. In order to measure the effectiveness of the metric, we collect human judgment scores for five translation system outputs. The test Chinese source sentence is randomly selected from our dataset, which contains 485 groups, and we choose first reference (given four for each source sentence) as standard reference. The SMT evaluation metric needs to score each translation system output against standard reference. Then we calculate the system correlation score and segment correlation score of metrics score following WMT metric share task standard. The result shown in the table 4 below

metric	segment	system
BLEU	.283	.512
Our method	.301	.528

Table 4: Our metric method compared with BLEU on Chinese English language pair.

Our method achieves better performance than BLEU. The segment score is measured following the method mentioned in Workshop on Statistical Machine Translation (Machacek and Bojar, 2014). The Pearson's correlation coefficient is used as the main measure of system-level metrics' quality. We measure the quality of metric's segment-level scores using Kendall's rank correlation coefficient as Vazquez-Alvarez and Huckvale suggested (2002).

However, the Pearson correlation of system-level performs poor in Chinese-English language pair as we shown in table 4, which means a huge disagreement in human judgment. We will collect more human judgment data to overcome this problem.

[2] LDC category numbers: LDC2000T50, LDC2002L27, LDC2003E07, LDC2003E14,LDC2004T07, LDC2005T06, LDC2005T10 and LDC2005T34.

7. Conclusion

In this work, we propose a bilingual discourse corpus annotation method. It is designed to improve the consistency between different languages. We annotated a bilingual discourse corpus in Chinese-English language pair, which contains nearly 20K sentences. We compare our annotated results with results generated by monolingual discourse annotation method in segment and structure. Our result is better than the traditional monolingual method. In order to explore the usage of our corpus, we design a simple translation metric based on BLEU by combining features extracted from bilingual discourse corpus. The result shows BLEU can benefit from our feature.

Acknowledgment

The research work has been partially funded by the Natural Science Foundation of China under Grant No. 61402478, the International Science & Technology Cooperation Program of China under Grant No. 2014DFA11350 and the West Light Foundation of Chinese Academy of Sciences, Grant No. LHXZ201301.

8. References

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In Proceedings of EMNLP , pages 334–343

William Mann and Sandra Thompson. 1988. Rhetor-ical structure theory: Toward a functional theory of text organization. Text , 8(3):243–281

Florian Wolf and Edward Gibsion. 2005. Representing Discourse Coherence: A Corpus-Based Study. Association for Computational Linguistics 31(2):249-287

Mei Tu, Yu Zhou, Chengqing Zong : Enhancing Grammatical Cohesion: Generating Transitional Expressions for SMT. ACL (1) 2014: 850-860

Shafiq Joty, Francisco Guzmán, Lluís M`arquez, and Preslav Nakov. 2014. DiscoTK: Using iscourse Structure for Machine Translation Evaluation. In Proceedings of the Ninth Workshop on Statistical Machine Translation

Daniel Alexandru Anechitei and Eugen Ignat , 2013. Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, Association for Computational Linguistics pages 72–76, Sofia, Bulgaria, August 9, 2013

Vanessa Wei Feng and Graeme Hirst, 2013. Detecting

Deceptive Opinions with Profile Compatibility. In Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP 2013), pages 338-346, Nagoya, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the ACL, pages 311-318

Matous Machacek and Ondrej Bojar, 2014. Results of the WMT14 Metrics Shared Task. Proceedings of the Ninth Workshop on Statistical Machine Translation , pages 293–301,Baltimore, Maryland USA, June 26–27,

Vazquez-Alvarez, Y. and Huckvale, M. (2002). The reliability of the itu-t p.85 standard for the evaluation of text-to-speech systems. In Hansen, J. H. L. and Pellom, B. L., editors, IN- TERSPEECH . ISCA.

Daniel Marcu, Lynn Carlson, Maki Watanbe,2000. The Automatic Translation of Discourse Structures. In Proc. of ACL 2000

Och FJ. 2000. GIZA++:Training of statistical translation models.(http://www-i6.informatik.rwth-aachen.de/~och/software/GIZA++.html)

Haoran Li, Jiajun Zhang, and Chengqing Zong. Predicting Implicit Discourse Relations with Purely Distributed Representations. In Proceedings of the joint conference of the Fourteenth China National Conference on Computational Linguistics (CCL) and the Third International Symposium on Natural Language Processing based on Naturally Annotated Big Data (NLP-NABD), Guangzhou, China, Nov. 13-14, 2015, pp.293-305

Mei Tu, Yu Zhou and Chengqing Zong. A Novel Translation Framework Based on Rhetorical Structure Theory. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics(ACL)(Short paper), Sofia, Bulgaria, August 4-9, 2013. Pages 370-374

Haitong Yang, Yu Zhou and Chengqing Zong. Bilingual Semantic Role Labeling Inference via Dual Decomposition. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Vol. 15, No. 3, Article 15, December 2015, 21 pages