Pisco Med Publishing

Original Research Article

Research on Process of the Bilingual Corpus Alignment Tool

Dandan Zhang

Shandong Jianzhu University Jinan, Shandong Province 250100

Abstract: As language researchers and translators gradually realize the importance of the development of corpus, many institutions at home and abroad have begun to devote themselves to the research and construction of corpus. During the process of corpus, bilingual corpus alignment is an indispensable step. However, the alignment software based on the existing alignment technology still can't meet the needs of users or translators. On the basis of previous studies, this paper mainly makes some beneficial attempts on the alignment process of sentence level automatic alignment technology in bilingual corpus. In this paper, Chinese and English files are imported into the corpus alignment tools and aligned one by one according to the translation units of sentence level. This paper presents the process of corpus alignment and proposes corresponding solutions to the errors in the process, to further improve the efficiency of bilingual corpus alignment.

Keywords: Bilingual corpus; Corpus alignment

1. Introduction

At present, translation is no longer a stagnant major. With the gradual recognition of the importance of corpus research and construction by scholars engaged in language research and machine single research, many institutions at home and abroad are committed to the research and construction of corpus.

Bilingual corpus alignment is an indispensable step in corpus construction. However, based on the existing alignment technology of alignment software, it still cannot fully meet the needs of users or translators. Therefore, how to better deal with the text and which alignment tool is suitable for different types of text to better complete the alignment task and improve the alignment efficiency is worth further study.

On the basis of previous studies, this paper mainly makes some beneficial attempts on the alignment process of sentence level automatic alignment technology in bilingual corpus. The source language and target language files of the text are imported into the corpus alignment tool, which is aligned one by one according to the translation units at the sentence level. By showing the problems in the process of corpus alignment and the process of solving the problems, the corresponding suggestions are provided for those who need the function of corpus alignment, and the best text processing method as well as the different types of text are obtained.

The alignment tools used in this article are ABBYY Aligner 2.0 and SnowCat.

1.1Corpus

Since the 1990s, many scholars have devoted themselves to the research and the development of corpus. Mona Baker and others have begun to apply corpus linguistics to translation studies. Baker (1995:230-236) believes that corpus can be divided into three categories: parallel corpus, comparable corpus and multilingual corpus[1].

From the linguistic point of view, corpus includes at least the following three points

(1) The corpus stores the language materials that actually appear in the process of using language;(2) Corpus is the basic resource of language knowledge on the basis of computer;(3) Real corpora need to be analyzed, processed to meet the need of useful research materials[2].

1.2 Bilingual Corpus

At present, many research institutions at home and abroad are committed to the construction of parallel corpora, and use these corpora to conduct in-depth research on various language phenomena.

In modern sense, bilingual corpus can be defined as the corpus in which the source text and the target text are mutually translation at the sentence level. The parallel corpus consists of two monolingual corpora. One corpus is the translation of another.

Bilingual corpus is a collection of original texts of one language and corresponding texts translated into another. Bilingual corpus can be used for contrastive study. By comparing the differences in vocabulary, sentence and style between the original text and the translated text, researchers can find the corresponding relationship in vocabulary and structure between the two texts[3].

1.3 Corpora Alignment

Corpus alignment generally refers to the association between the corresponding segmented segments in the parallel corpus of bilingual texts, which can be defined from different perspectives according to the specific content of the corpus.

Given a text and its translation, an alignment is a segmentation of two texts such that the nth segment of one text is the translation

doi: 10.18282/l-e.v10i5.2664

of the nth segment of the other (Simard, Foster, & Isabelle,1992). Empty segments are allowed which can be corresponding either to translator's omissions or to additions. In other words, alignment is the process of finding relations between a pair of parallel documents. An alignment may also constitute the basis of deeper automatic analyses of translations[4].

There are various angles of corpus alignment, which indicates that the classification of corpus alignment is also diversified. When Huang Junhong reviewed the foreign corpus alignment techniques, they divided them into four categories:

(1) sentence level alignment, (2) Lexical alignment, (3) Unit alignment of multi word combination, which means the collocation alignment of phrases or words. (4) Clause and paragraph alignment[5].

1.4 Sentence alignment tool

There are several documented algorithms and tools for sentence level alignment. Generally, they can be divided into three categories: Based on length, based on dictionary, or based on partial similarity.

Generally, sentence aligners take as input the texts to align, and, in some cases, additional information, such as a dictionary, to help establish the correspondences[6].

A typical sentence alignment algorithm starts by calculating alignment scores, trying to find the most reliable initial points of alignment – denominated "anchor points". This score may be calculated based on the similarity in terms of length, words, lexicon or even syntax-tree [Tiedemann 2010][7].

2. Research process

The research process of this paper includes

1.First alignment result.2.Problems of the first results.3.Suggestions.4.Second alignment results.5.Result analysis.

2.1 Corpora Acquiring

In order to make the paper more practical, and the conclusions, suggestions and the applicability of the method is widened, author selects *The Speech by Chinese President Xi Jinping At the Opening Session of the World Economic Forum Annual Meeting 2017* (hereinafter referred to as Speech), and *Code for Design of Building Foundations GB 5007-2002* (hereinafter referred to as GB 5007-2002).

The first material is a typical government report. Both Chinese and English versions are very standard. Therefore, the layout of this material is exquisite, detailed, and appropriate, and the language expression should be clear and concise. The language expression with Chinese characteristics is very suitable for alignment material. In addition, the differences between Chinese and English expressions can be clearly seen in the government report, which is an indispensable material for translation learners. There are obvious differences between sentences of Chinese and English, which can reflect more problems in the alignment process, so it is more suitable for alignment corpora.

Secondly, considering that the construction of corpus and research are very popular, and as a very important part of it, corpus alignment is the focus of each process. Therefore, in addition to the corpus of government reports, some practical and official guidance materials are also needed. The second article "GB 5007-2002" is revised by China Academy of Building Sciences. This specification is divided into 10 chapters and 22 appendixes. This material is a typical standard specification.

The reasons for selecting this article for alignment are as follows:

(1)This article has various format.(2)The content includes words, numbers, tables, formulas, etc.3.Strong practicability

In conclusion, the above two materials are suitable for alignment. After confirming the corpus, this paper will show the alignment process of each alignment tool.

2.2 The processes of alignment-SnowCat

Considering the length of the second article is too long, it is not suitable to align all of them. Therefore, after reading the whole article, choose the third chapter that covers all the contents mentioned above as the target of alignment material.

2.2.1 First alignment result:

Speech

File name	Number of words		Number of sentences		
	Original text-E	Target text-C	Original text-E	Target text-C	Sentence pairs
Speech First alignment	4356	6678	165	163	163

GB 5007-2002.

File name	Number of words		number of sentences		
	Original text-E	Target text-C	Original text-E	Target text-C	Sentence pairs
GB 5007-2002 Chapter 3 First alignment	2144	2524	154	148	146

2.2.2 Problems of first result

The Speech:

(1)The original text does not match the number of target text, appearing blank lines.2.Can't divide sentences according to punctuation.3.Wrong lines appear from the beginning of the third sentence, and then the Chinese and the English are staggered.4.Although after Arabic numerals appear in Chinese sentences, which could help the work of alignment, the original text is aligned with the target text, it is not guaranteed that the sentences around the sentence with numbers is aligned.5.The overall alignment effect is poor, and a lot of manual changes are needed in the later work.

The Chapter 3 of *GB 5007-2002*.:

(2)The original text does not match the number of interrogative sentences, appearing blank lines.2.Can't divide sentences

according to punctuation.3.Even if both the original and the target text have numbers, they still can't be aligned.4.Invalid number to number result in table alignment.5.The whole English sentence is split.6.Formulas affect alignment results.7.The overall alignment effect is poor, and a lot of manual correction is needed in the later stage.

2.2.3 Suggestions:

1.Turn off the revision function and accept the revision mark2.Cancel auto numbering3.Replace page breaks and splitters with blanks4.Manually divide the article into several parts, insert numbers where you can quickly distinguish between Chinese and English, and form separate lines.5.Remove header and footers6.Forms:

(1)Converts a table to paragraph markers in text

(2)Delete the unimportant numbers, formulas, etc. in the table (because it is not needed to align the numbers and formulas, the construction of corpus is not affected)

2.2.4 Second alignment result

File name	Number of words		number of sentences		
	Original text-E	Target text-C	Original text-E	Target text-C	Sentence pairs
Speech Second alignment	4360	6682	169	167	167

File name	Number of words		number of sentences		
	Original text-E	Target text-C	Original text-E	Target text-C	Sentence pairs
GB 5007-2002 Chapter 3 Second alignment	1903	2341	71	70	146

2.2.5 Result analysis

First material:

However, there are still some problems that the original text and target text cannot be segmented according to punctuation, so it is suggested to reprocess manually.

Improvement after taking the suggestions-- *GB 5007-2002*:

1.The number of alignment lines of the original text is the same as that of the target text.2.No major errors.3.Invalid alignment is significantly reduced.4.Alignment effect is good, saving a lot of labor and time, improving efficiency, so it is recommended to use the suggestions.

2.2.6 Summary

The following is a summary of scat alignment results

File name	Number of words		number of sentences		
	Original text-E	Target text-C	Original text-E	Target text-C	Sentence pairs
Speech First alignment	4356	6678	165	163	163
Speech Second alignment	4360	6682	169	167	167
GB 5007-2002 Chapter 3 First alignment	2144	2524	154	148	146
GB 5007-2002 Chapter 3 Second alignment	1903	2341	71	70	146

2.3 The processes of alignment- ABBYY Aligner 2.0

2.3.1 First alignment result:

File name	Number of words		Sentence pairs	Blank lines
	Original text-C	Target text-E		
Speech First alignment	6678	4356	206	30
GB 5007-2002 Chapter 3 First alignment	2551	2118	221	30

Sentence splitting is better than SCAT, but due to the different sentence patterns and expressions between Chinese and English, some English sentences are divided into clauses in Chinese, which leads to the blanks in Chinese.

The first alignment of *Speech* by ABBYY Aligner 2.0 is better than SCAT. The alignment effect is good. **The subsequent work is to delete the blank line, no second alignment is necessary for this file.**

2.3.2 Problems of first result

Speech:

1.The original text does not match the number of target text, appearing blank lines:2.Can't divide sentences according to punctuation:3. Wrong lines appear from the beginning of the third sentence, and then the Chinese and the English are staggered.4. Although after Arabic numerals appear in Chinese sentences, which could help the work of alignment, the original text is aligned with the target text, it is not guaranteed that the sentences around the sentence with numbers is aligned.5. The overall alignment effect is poor, and a lot of manual changes are needed in the later work.

GB 5007-2002:

1.The original text does not match the number of interrogative sentences, appearing blank lines.2.Can't divide sentences according to punctuation.3.Even if both the original and the target text have numbers, they still can't be aligned.4.Invalid number to number

result in table alignment.5.The whole English sentence is split.6.Formulas affect alignment results.7.The overall alignment effect is not good, and a lot of manual correction is needed in the later stage.8.Both the original text and the target text appear one to zero or zero to one.9.Chinese sentences, numbers and publicity are disconnected for no reason.10.Empty lines appear in alignment results.11. The alignment effect of some numbers is also poor.12.Many invalid alignments: punctuations to punctuation, number to number.

2.3.3 Suggestions

1.Turn off the revision function and accept the revision mark2.Cancel auto numbering.3.Replace page breaks and splitters with blanks.4.Replace empty lines with empty ones.5.Remove header and footers.6.Delete the unimportant numbers, formulas, etc. in the table (because the numbers and formulas are not words, no translation is needed, and the construction of corpus is not affected).7.Converts a table to paragraph markers in text.8.After the above steps are completed, copy the full text to a new document and paste only the text, deleting format.

3 Conclusion

The research and development of corpus is on the right track, which is a hot topic discussed by scholars in recent years. The basis of machine translation is high-quality parallel corpus. Only by solving the construction of bilingual corpus, can machine translation help people better and more effectively.

This paper is a useful attempt to the alignment process of sentence level automatic alignment tool in bilingual corpus. Whether scholars or enterprises are creating a corpus, the first step is alignment. Therefore, bilingual alignment is a crucial part of corpus construction. This paper attempts to study the whole process of corpus alignment: first alignment, suggestions for improvement, second alignment, and alignment result analysis. A series of attempts during the whole process of are summed up as follows,

3.1Comparison of SCAT and ABBYY Aligner

The drawback of SCAT is that English can't automatically divide sentences according to the need of alignment

The disadvantage of ABBYY Aligner is that continuous numbers are easy to separate the whole sentence, which greatly increases the workload of manual work

SCAT is more suitable for aligning the second material--*GB 5007-2002*, which contains more tables, formulas and numbers.

ABBYY Aligner is more suitable for aligning the first file--*Speech*, which is the kind of material full of words, and the alignment effect is better.

Pretreatment of original text:

In the process of alignment, after importing the original text directly into the alignment tool, the alignment effect is not ideal, and it needs a lot of labor and time. Therefore, this paper makes a series of attempts and puts forward some suggestions on the revision of the text. After preprocessing the alignment material, the alignment effect can be greatly improved, saving time and labor.

3.2 The deficiency of the paper

Due to the limited time and experience, this paper can only do some superficial attempts, there are still many works to be further completed, such as:

The alignment materials do not involve literature and other subjects, so the suggestions put forward in this paper are only suitable for government work report materials and large-scale articles involving figures, tables, etc.

This paper only deals with SCAT and ABBYY Aligner alignment tools, and the scope of practice is small

At present, corpus research and construction are in full swing, computer-aided translation technology continues to meet the needs of people's translation, but there are still a variety of problems in bilingual alignment technology and corpus construction, which need researchers of various disciplines to continue to work hard and contribute their own strength.

References:

[1] Baker M. 1995. Corpus in translation studies：An overview and some suggestions for future research [M]. *Target*: 230-236.

[2] Zhao Xiaoman. 2010 sentence-Level Alignment of English-Chinese Parallel corpus and its Application in Machine Translation [C]. Anhui University: 1.

[3] Christopher C. Yang *, Kar Wing Li. 2003. Building parallel corpora by automatic title alignment using length-based and text-based approaches [J]. *Information Processing & Management*: 2.

[4] Simard, M., Foster, G., & Isabelle, P. 1992. Using cognates to align sentences in bilingual corpora [J]. In Fourth international conference on theoretical and methodological issues in machine translation (TMI-92), Montreal, Canada: 1072.

[5] André Santos. 2011. A survey on parallel corpora alignment [J]. *MI-STAR*: 122.

[6] Tiedemann, J. 2010. Lingua-Align: An Experimental Toolbox for Automatic Tree-to-Tree Alignment[J]. Department of Linguistics and Philology Uppsala University, Uppsala/Sweden: 742.