Research on Alignment in the Construction of Parallel Corpus

Zhaorong Zong\textsuperscript{1,}\textsuperscript{a*}, Changchun Hong\textsuperscript{1,}\textsuperscript{b}

\textsuperscript{1}School of Foreign Languages, Huangshan University, 39 Xihai Road, Tunxi District, Anhui Province, China
Email: \textsuperscript{a} zongzhaorong@hsu.edu.cn, \textsuperscript{b} hongcc@hsu.edu.cn
\* Corresponding author

Abstract: Parallel corpora are of great value in the field of machine translation and cross-language information retrieval. Benefiting from the development of machine learning and deep learning, the technology of the construction of corpus evolves from vocabulary alignment, phrase alignment to chunk alignment. The high quality of automatic bilingual chunks alignment in corpus plays an important role in the performance improvement of machine translation systems, especially in computer-aided translation systems. In the study, the degree of adhesion and relaxation is used to measure the tightness and looseness of the inter-word connection when a chunk is identified, which can be expressed by a mathematical model. The task of chunk alignment in the construction of a parallel corpus can be described as the three steps: input bilingual sentences, segment chunks, and semantic alignment. At present, most algorithms are based on statistical methods, and the output alignment results are machine-oriented.

1. Parallel Corpus

With the rapid development of computer technology, communication technology and artificial intelligence, the world is getting smaller and smaller, and the communication between people is getting closer. In the process of communication, the language barrier problem is highlighted, which directly promotes the rise of cross-language information processing technology. Benefiting from the development of machine learning and deep learning, the corpus plays an increasingly important role as a basic resource. Corpus refers to the collection of original corpus with certain structure, certain representativeness and certain scale for specific application purposes \textsuperscript{[1]}. The current corpus is usually recorded on a computer and can be retrieved by a computer program. \textsuperscript{[2]} The multilingual corpus, currently composed of two languages, is the most common, also known as the bilingual corpus. According to the difference of mutual translation, the bilingual corpus is divided into parallel corpus and comparable corpus. Parallel corpus is a collection of text pairs consisting of source language text and translated target language text and there is a strict translation relationship between the two languages. Comparable corpus is a non-translated text pair collection with different language but similar content.

At present, the construction of corpus is based on algorithm system whose core technology is a deep neural network with massive nodes. \textsuperscript{[3]} As a result, a super-size corpus or even a hyper-size corpus is needed. Table 1 shows the differences between traditional corpus and AI corpus in capacity and time span.

| Corpus Name                        | Words       | Language | Time period   |
|------------------------------------|-------------|----------|---------------|
| Global Web-Based English(GloWbE)  | 1.9 billion | 20 countries | 2012-2013     |
| Corpus of Contemporary American English(COCA) | 450 million | American | 1990-2012     |
| Corpus of Historical American English(COHA) | 400 million | American | 1810-2009     |
| TIME Magazine Corpus               | 100 million | American | 1923-2006     |
| Corpus of American Soap Operas     | 100 million | American | 2001-2012     |
Parallel corpus has shown great value in the field of machine translation and cross-language information retrieval, because it contains a large amount of bilingual knowledge as a basic resource for cross-language information processing. Labov believes that "Every major paradigm shift in the field of linguistics is triggered by the changes in the perception of basic data in the subject area (Stubbs, 1993: 24). However, with the development of computer technology and the improvement of corpus construction, larger-scale corpus needs to be processed and dealt with. The corpus research is developing into a coherent, comprehensive, and rich paradigm, which is embarking on a series of issues related to translation theory, description, and practice (Laviosa, 1998:474).[4] The parallel corpus contains two different levels of corresponding languages in words, phrases, sentences, paragraphs, and chapters, providing a research foundation for cross-language information processing technology, which has long attracted the attention of scholars.

2. Acquisition of natural corpus

In the early stage of parallel corpus construction, the collection and processing of corpus mainly relied on manual participation in the selection and sorting. Its sources were mainly international and domestic conference records, religious works, literary and artistic works and product specifications. This kind of acquisition greatly limited the construction efficiency of the parallel corpus, restricting its expansion in the scale and field, and it is difficult to meet the requirements of timeliness.

With the development of the network, more and more websites have begun to provide more than two language versions to meet the needs of their users. More and more online information is being distributed in multiple languages, making different websites or the same webpage flooded with a large number of multilingual resources, which provides a solid data foundation for the acquisition of bilingual resources based on web mining.[5]

It is a common method of studying language by using computers to collect, assemble, and process lingual data to form a corpus. The corpus is one of the key resources for natural language processing, and its research began in the 1970s. So far, a large number of corpora have been built in the world, such as BROWN, LOB, COBUILD, LONGMAN, BNC and ICE.[6]

Parallel corpus is a very useful resource for language studying. For example, in the field of statistical machine translation, it is necessary to use a parallel corpus training model to perform translation work. The result of translation is virtually decided by the size of the corpus. In general, the larger the corpus is, the higher the translation quality is[7]. However, the initial acquired bilingual corpus contains a lot of noise, and the missing, misplaced information needs to be processed before they can be used. (Fig 1) The above problem can be solved by using the corpus alignment technique. The corpus alignment technique finds the most matching lingual chunk information according to the characteristics of the two languages. The characteristic chunk features include length information, vocabulary information, punctuation and position information. Accurate analysis of the feature information of the bilingual chunk can improve the accuracy of the alignment and obtain high quality of a parallel corpus.

![Fig 1. noisy channel model](image)

Although the corpus can provide convenient and fast data retrieval for research, corpus construction itself is a
time-consuming and laborious task from the collection of corpus (including scanning, identification, proofreading, etc.), pre-processing, to corpus partition, label, alignment, and the workload of each step is massive, especially the corpus alignment work, which needs to be done as precisely as possible to ensure the accuracy of the later corpus processing and future retrieval.

3. The corpus alignment

The alignment of corpus refers to storing the text of the source language and the corresponding translation text, and aligning the two texts at different language levels (such as chapters, paragraphs, sentences, phrases, words, etc.). Corresponding units is an important concept in the construction of a parallel corpus, which refers to any identifiable corresponding fragments in the source text and translation text in the parallel corpus. It is not just a theoretical concept, but also an operational concept. According to certain working principles, if the processor of the text considers that two bilingual segments in the parallel corpus are corresponding in the translation process, it can be identified and separated.

Alignment of vocabulary and phrases is relatively mature nowadays. The current research focus is on the partition and alignment of chunks, which is of great significance in language research. However, the bilingual correspondence at the level of the sentence has a small recurrence rate, while the computer-based translation of machine-assisted translation is inseparable from frequency information. Without a high rate of recurrence, it is difficult to obtain meaningful frequency information, and the significance of machine-assisted translation research becomes meaningless.

1) Alignment of vocabulary

The corpus alignment can be based on the lexical level, which corresponds to the vocabulary of the source language and its corresponding translated language. However, in addition to the one-to-one corresponding segment at the vocabulary level, there are also other forms, such as one to several, one to nil, etc. As a matter of fact, the vocabulary in the form of correspondence is very complicated, and the researchers can use the words prefabricated in the background to match some of the sentence segment one by one. With computer aiding, the bilingual segments can be paired in a one-to-many or many-to-one relationship. In translation practice, the computer can display one-to-many translations to the translator for selection or reference according to the frequency. The problem is that pure frequency information is obviously not enough for translation, for some specific lexical meaning cannot be separated from the context. That is to say, without the context, the meaning of the vocabulary cannot be completely determined.

2) Alignment of phrases

The study of phrase units in corpus linguistics generally adopts a corpus-driven research method, using concept definition → database construction → work method confirmation → multi-word structure extraction → linguistic analysis (grammatical level, semantic level, pragmatic level) → function clustering (conceptual function, textual function, interpersonal function). In the past, the traditional bilingual text alignment was based on the method of “analysis-analysis-matching”. First, syntactic analysis or multi-layer syntax analysis was performed on sentences of a certain language text, and a heuristic process was utilized to align each pair of sentences and phrases in two languages to obtain a phrase-aligned bilingual system. Using the corpus method for alignment and machine translation training, the statistical information of the co-occurrence frequency is applied to align phrases, that is, the corresponding information is found through the co-occurrence of the phrases in the entire corpus.

At present, the ordering model in phrase statistical machine translation can be divided into two categories: one is a non-hierarchical phrase ordering model that does not use lexical information for simple shifting models and flat ordering models, and a variety of lexical phrase ordering models; the other is a hierarchical phrase ordering model, which is based on a tree-like syntactic hierarchy and is adjusted layer by layer. Since the syntactic hierarchy can cover different ranges of phrases from large to small, the introduction of syntactic hierarchy helps to solve the problem of long-distance scheduling in statistical machine translation.

With the continuous maturity of natural language processing technology, the research on corpus-driven phrase unit has alienated from the field of single language research. In general, the phrase-based statistical machine translation system considers all the phrase segments of a sentence to be uniformly distributed, which is obviously unreadable because the phrases are not as easy to define as segments of words. Therefore, the method of chunk alignment is introduced in corpus construction.

3) Alignment of chunks

The concept of chunk was first proposed by Abney, representing the non-recursive core components of the sentence. The terminology used in the lingual chunk research is very complicated, including “chunks”, “lexical chunks”, “prefabricated chunks”, “prefabricated patterns”, “formulaic languages”, “speech formula”, “formulaic
unit”, “collocation”, “lexical phrase”, “ready-made complex unit”, “lexicalized sentence stem”, etc. These different terminologies reflect both the importance of the study of “chunk” and the complexity of it. What they have in common is that they emphasize that language users identify chunks from context rather than from rules.

The task of chunk alignment in the construction of a parallel corpus can be described as the three steps: input bilingual sentences, segment chunks, and semantic alignment. At present, most algorithms are based on statistical methods, and the output alignment results are machine-oriented and cannot be interpreted manually. High quality chunk alignment is of high value for the construction of a bilingual parallel corpus. In phrase-based machine translation systems, long sentences are often discarded because of the time-consuming work. Using high-quality alignment chunks to “disassemble” long sentences into shorter inter-translated segments can reduce the time used and make full use of the corpus.

In the source language partition, the degree of adhesion and relaxation are used to measure the tightness and looseness of the inter-word connection when the chunk is identified, which can be expressed by mathematical mode as follows:

Given a sentence consisting of N words $S = w_1 w_2 w_3 \ldots w_i \ldots w_{n-1} w_n$, the gap between adjacent words $w_i, w_{i+1}$ is recorded as a separating point $g_i$, whose values $1, 0$ respectively represent whether the chunk is separated. Thus, we can use the sequence of values of the partition points $G = g_1 g_2 g_3 \ldots g_i \ldots g_{n-2} g_{n-1}$ to represent the chunk division. Set the two properties of adhesion and relaxation for each partition point $g_i$ (all values are positive integers): the degree of adhesion represents the tightness of the connection between $w_i, w_{i+1}$, which is recorded as $a_i$, and the greater the value, the closer the connection is; conversely, the degree of relaxation represents the looseness of the connection between adjacent words $w_i, w_{i+1}$, which is denoted as $r_i$, and the larger the value, the less tight the connection is. Assuming the degree of adhesion $a_i$ and relaxation $r_i$ between $w_i, w_{i+1}$ are known, the probability that the chunk division should be separate at the partition point $g_i$ is approximate to the formula (1).

$$P(g_i = 1) \approx \frac{r_i}{a_i + r_i} \quad (1)$$

If the partition point $g_i$ should not be separate, the probability that $w_i, w_{i+1}$ is divided into the same chunk is approximate to the formula (2).

$$P(g_i = 0) \approx \frac{a_i}{a_i + r_i} = 1 - P(g_i = 1) \quad (2)$$

Both the degree of adhesion and relaxation are calculated independently and separately, and different calculation models have different demands. After the sentence is divided into chunks, several chunks are formed. For a chunk containing M words $C = w_{i+1} w_{i+2} \ldots w_{i+m} \ldots w_{i+M-1} w_{i+M}$, its left and right partition points $g_i, g_{i+M}$ are taken a value of 1, while all internal partition points a value of zero. Since each feature model can give the degree of adhesion and relaxation of each word a partition point $g_i$, which constitutes a huge state space, and each feature weight needs to be optimized and then merged.[11]

4. Conclusion
The study on corpus alignment of parallel corpus is an important part of constructing a corpus. As natural language processing techniques mature, the research on corpus-driven phrase unit has evolved from vocabulary alignment, phrase alignment to chunk alignment. The high quality of automatic alignment bilingual chunks in corpus plays an important role in the performance improvement of machine translation systems, especially in computer-aided translation systems. In the next step, the algorithm of chunk alignment mode will be further studied and the scope of research will be expanded to form a complete chunk alignment mode and provide technical support and corpus for parallel corpus construction.

Acknowledgements
This paper is financially supported by the following:
Anhui Social Sciences Key Project Funding SK2016A0878
Anhui Social Sciences Project Funding SK2015A812

References:
[1] Wray, A. Formulaic Language and the Lexicon[M]. Cambridge: Cambridge University Press, 2002.
[2] Hong, Changchun Research on the Information Processing Model of Source Language of Parallel Corpus[C]//1st Conference on Advanced Algorithms and Controlling Engineering, ICAACE2108, 10-12
[3] Franz Josef Och. Statistical Machine Translation: From Single-Word Models to Alignment Templates[ D] . Germany : Computer Science Department ,RWTH Aachen, October , 2002.
David Chiang . A hierarchical phrase-based model for statistical machine translation [A]. In Proceedings of ACL 2005 [ C] , page 263-270.

[4] Wray, Alison. Formulaic language: Pushing the boundaries [M] Oxford: Oxford University Press. 2008: 9.

[5] Marina D, Chi W. Automatic Construction and Ranking of Topic Key-phrases on Collections of Short Documents[C]//Proceeding of 2014 SIAM International Conference on Data Mining. Philadelphia, Pennsylvania: SIAM, 2014: 398-406.

[6] Tiedemann. J. Bitext Alignment. San Rafael: Morgan & Claypool Publishers. 2011. 106 Liu J, Shang J, Wang C, et al. Mining Quality Phrases from Massive Text Corpora[C]//Proceedings of the 36th International Conference on Management of Data, Australia:ACM,2015:1729-1744.

[7] Zong, Zhaorong Research on the Relations between Machine Translation and Human Translation,[C]//1st Conference on Advanced Algorithms and Controlling Engineering, ICAACE2108, 10-12 Aug., 2018,Chongqing Volume 1087,062046

[8] Cohen J D.Highlights: language and domain-independent automatic terms for abstracting[J]. Journal of the American Society for Information Science,1995, 46(3):162-174.

[9] Chuang , T.C.& K.C.Yeh. Aligning parallel bilingual corpora statistically with punctuation criteria. International Journal of Computational Linguistics & Chinese Language Processing, 2005 ( 1) : 95-122

[10] Abney Steven. “Statistical Methods and Linguistics.” [J]. The balancing act: Combining symbolic and statistical approaches to language. 1996:1-26

[11] Yu Jingsong, Wang Huilin, Wu Shenglan A Bilingual Chunk Alignment Algorithm For Computer-aided Translation[J].Journal of Chinese Information Processing, 2015,29(01):67-74.