ISSN 2616-7433 Vol. 5, Issue 11: 55-59, DOI: 10.25236/FSST.2023.051109

Key Technologies for Constructing Bilingual Corpus for English-Chinese Translation

Yanming Li*

School of Applied Technology, University of Science and Technology Liaoning, Anshan, China
110017218@qq.com
**Corresponding author*

Abstract: *Bilingual corpus provides translators with vivid language resources and translation examples. Bilingual corpus can extract the original text and the corresponding translation at the same time, take translation based on parallel corpus as a method to cross the language boundary between source language and target language, and use bilingual dictionary as the main knowledge source to realize query translation processing. Based on the relevant research results of corpus linguistics, this paper studies the key technologies of constructing bilingual corpus for English and Chinese translation. Among them, corpus collection, based on the basic principle of web crawler technology, designs the corpus collection process, which provides the basis for system programming. Corpus alignment finds the corresponding source text and target text in bilingual or multilingual texts, and recommend using ABBYY Aligner tool. Corpus annotation studies the classification and basic principles of annotation and solves the core problems of annotation.*

Keywords: *English-Chinese Translation; Bilingual Corpus; Key Technologies; Corpus collection; Corpus alignment; Corpus annotation*

1. Introduction

Since the mid-1990s, with social development and economic progress, research related to corpus has developed rapidly at home and abroad, and corpus has gradually become an important research direction in the field of linguistics [1]. According to the number of languages contained in the corpus, it can be divided into monolingual corpus, bilingual corpus and multilingual corpus. Among them, bilingual corpus and multilingual corpus can be divided into parallel corpus and comparative corpus according to the organization form of corpus. The corpora collected in parallel corpora form a translation relationship with each other, which is mostly used in bilingual dictionary compilation, teaching Chinese as a foreign language and machine translation. A comparative corpus is a collection of texts in different languages with exactly the same description, which are only expressed in different languages, and is used for language comparative research to investigate the differences between different kinds of natural languages. In addition to its diverse and rich advantages, translation corpus can also create a real lexical ecological context [2]. Corpus and computer-assisted translation are increasingly important in translation teaching and practice [3]. Translation corpus can help translators determine the semantic rhyme of the original words, the semantic rhyme of the author and the original style, obtain reference words, sentences and expression methods, and quantitatively evaluate the translation, which breaks through the limitations of traditional translation and becomes an effective means to solve translation errors. It can also explore the process, characteristics and laws of two languages and their transformation, and improve the efficiency of translation teaching.

2. Corpus Collection

Corpus collection is the most complicated work in corpus construction, which requires a lot of manpower and material resources. In order to ensure the authenticity and vividness of the corpus, the construction of bilingual corpus for English translation needs to focus on the materials accumulated by English teachers every day, and widely use technologies such as web crawler to collect corpus through multiple channels.

ISSN 2616-7433 Vol. 5, Issue 11: 55-59, DOI: 10.25236/FSST.2023.051109

2.1 Basic Principles of Web Crawler Technology

Web crawler is a program that automatically extracts web pages, downloads web pages from the World Wide Web for search engines, and is an important component of search engines [4]. The traditional crawler starts with the URL of one or several initial web pages and obtains the URL on the initial web page. In the process of crawling the web page, it continuously extracts new URLs from the current page and puts them in the queue until it meets certain stopping conditions of the system. The workflow of focused crawler is complex, so it is necessary to filter links irrelevant to the topic according to a certain web page analysis algorithm, keep useful links and put them in the URL queue waiting to be crawled. Then, according to a certain search strategy, the URL of the next page to be crawled will be selected from the queue, and the above process will be repeated until a certain condition of the system is reached. In addition, all the web pages crawled by the crawler will be stored by the system, analyzed and filtered to some extent, and an index will be established for later query and retrieval; For the focused crawler, the analysis results obtained in this process may also give feedback and guidance to the subsequent crawling process. The basic principle of web crawler technology is shown in Figure 1 [5].

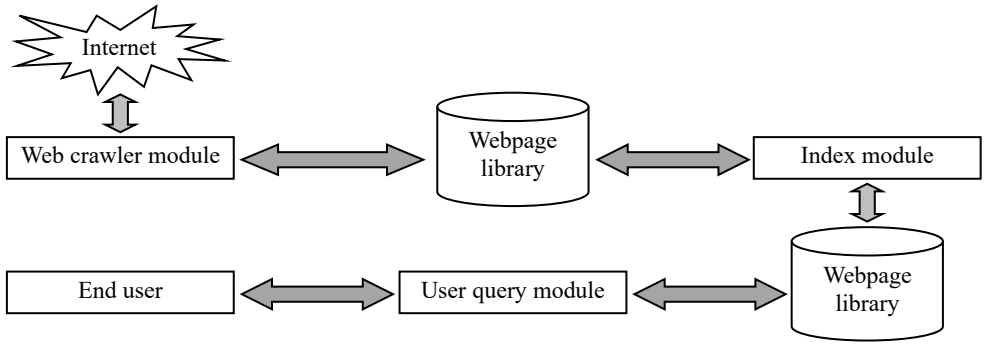

Figure 1: Basic principles of Web crawler technology

2.2 Corpus Collection Process Based on Web Crawler

The corpus collection process based on web crawler is shown in Figure 2 [6].

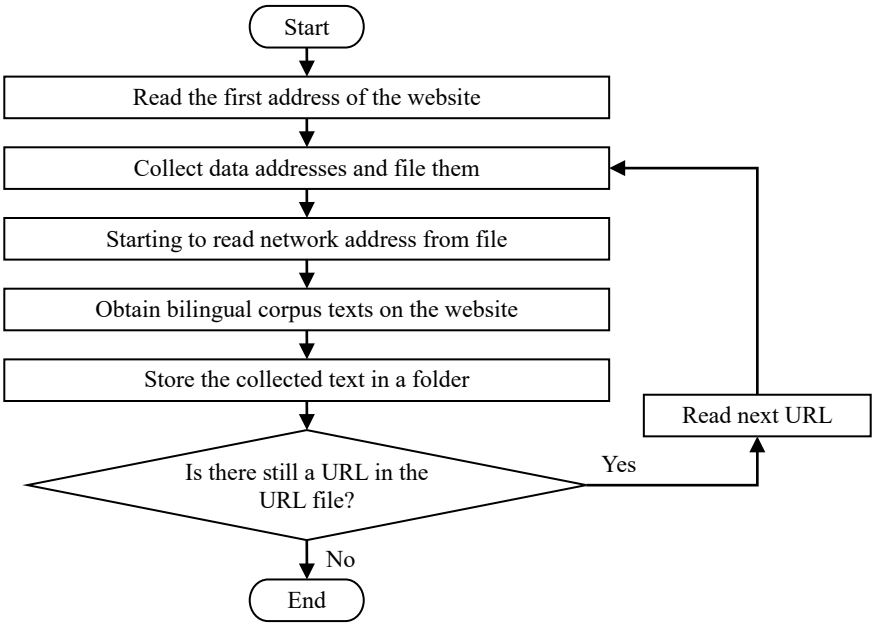

Figure 2: The process of collecting corpus based on web crawlers

ISSN 2616-7433 Vol. 5, Issue 11: 55-59, DOI: 10.25236/FSST.2023.051109

3. Corpus Alignment

Corpus alignment is the key technology to establish translation memory, that is, to find the corresponding source text and target text in bilingual or multilingual texts [7]. After the completion of a translation project, the translator or translation company should make the original text and translation into bilingual or multilingual aligned corpus, regularly update and maintain the translation memory, and properly preserve it as a language asset. When encountering a translation project in the same or similar field, the corresponding translation memory can be retrieved, so as to save manpower and material resources and improve the quality and efficiency of the translation project. Corpus alignment plays an important role not only for translators and translation companies, but also for cultural exchange and knowledge dissemination.

Corpus alignment refers to the establishment of correspondence between different language units of two or more language texts, that is, to determine which language unit of the source text and which language unit of the target text are mutually translated. According to the size of language units, corpus alignment can be divided into lexical level, sentence level and paragraph level. Different application purposes need different levels of alignment, for example, lexicography needs lexical level alignment. The construction of translation corpus needs sentence-level alignment. For bilingual texts with a high degree of variation, it is difficult to achieve lexical and sentence-level alignment, so paragraph-level alignment is needed. On the whole, the current corpus alignment mainly focuses on lexical alignment and sentence alignment. Sentence-level alignment is a more commonly used alignment mode in translation practice, and its working principles mainly include the following three types [8].

(1) Length-based alignment method. According to the length ratio of the two languages, the bilingual text is aligned at the sentence level. Some scholars put forward a sentence alignment method based on punctuation marks. Through the alignment experiment of Chinese-English parallel corpus, it is found that this algorithm is superior to the length-based alignment algorithm. Under the common statistical framework, using both punctuation-based and length-based algorithms can effectively improve the accuracy of Chinese-English bilingual text alignment.

(2) Lexical-based alignment method. Find as many corresponding bilingual words as possible in double sentence pairs, and align the sentence pairs with the largest corresponding number. On this basis, some scholars have constructed a one-to-one statistical translation model of vocabulary, which improves the accuracy of the alignment algorithm. Because Chinese and English belong to two language families, it is impossible to use cognate word positioning method to improve the alignment accuracy. Therefore, although the alignment method based on cognates is effective in English-French bilingual alignment, it is not equally effective in Chinese-English alignment.

(3) Hybrid alignment method. The length-based method is fast, but it is not robust, and it is easy to cause continuous misalignment. If the bilingual text is omitted, or the proportion of words in the bilingual text deviates due to different translation styles, the spread of error alignment will occur. The method based solely on vocabulary has good robustness, and it can achieve rapid positioning when dealing with some highly recognizable language units. Although this method is accurate, it is time-consuming and difficult to meet the technical requirements of large-scale corpus construction. Combining the length-based method with the lexical-based method, the alignment experiment of Chinese-English text and Japanese-English text is carried out, which improves the accuracy of bilingual alignment. Therefore, at present, most bilingual alignment tools combine these two methods to achieve automatic bilingual alignment.

4. Corpus Annotation

An important sign of the biggest difference between corpus and paper materials is machine-readable. In order to give full play to the characteristics of fast computing speed and strong computing power of computers, the corpus must be marked in advance. The process of corpus annotation is a formalization of language knowledge. The quality and depth of corpus annotation directly affect the richness and accuracy of information that can be excavated from corpus, which determines the availability and value of corpus. The disambiguation model is established by using the semantic annotation corpus and applied to the machine translation system, which effectively improves the translation effect [9].

ISSN 2616-7433 Vol. 5, Issue 11: 55-59, DOI: 10.25236/FSST.2023.051109

4.1 On the Classification of Corpus Annotation

Annotation is a process of expressing implicit knowledge form, an important link of transforming unstructured text into semi-structured text and a process of transforming text into knowledge. Corpus annotation is usually divided into the following three categories.

(1) Part of speech annotation. Part-of-speech annotation is to give each word in the text a corresponding part-of-speech tag, including punctuation marks. Part of speech markers represent the grammatical features of a word, so they are also called grammatical markers. The general process of automatic part-of-speech annotation includes preprocessing the text to be tagged, automatic word segmentation, tagging all the words in the text, and eliminating ambiguous codes. Part-of-speech annotation is of great significance, providing materials for higher-level natural language text processing, providing detailed information for linguistic research and obtaining the knowledge of part-of-speech annotation of parts of speech and frequency from the processed text.

(2) Semantic annotation. Semantic annotation mainly includes semantic features of language units and semantic relations between language units. The so-called automatic semantic tagging of words means that the computer uses logical operation and reasoning mechanism to correctly judge and label the meanings of words that appear in a certain context. The step of semantic tagging is that each polysemous word that needs to be tagged with meanings can be clearly distinguished in advance. For each polysemous word that appears in a specific context, determine an appropriate meaning.

(3) Phonetic/prosodic annotation. Phonetic annotation marks the changes of phonemes in syllables in the language stream. Prosodic annotation system is a set of machine-readable phonetic prosodic transliteration symbols and rules. ToBI (Tones and Break Indices) is an internationally accepted prosodic tagging system for spoken corpus, which was first designed as a prosodic tagging standard for standard American English from 1991 to 1994. ToBI's design principle is to label the distinctive intonation patterns and prosodic units in the language. ToBI's design criteria include four aspects. Reliability, that is, the consistency of annotation results of different annotators should reach at least 80%; Comprehensive, that is, covering the most important prosodic phenomena in natural language; Easy to learn, that is, to learn in a short time; Compatibility means combining with the latest methods of speech synthesis and speech recognition and the current theories of syntax, semantics and pragmatics.

4.2 Principles of Corpus Annotation

At present, the annotation types of large corpora in the world are very similar, and they follow the internationally accepted corpus annotation principle, which is the seven basic principles of corpus annotation proposed by Leech [10].

(1) Annotations and attached codes can be deleted and restored to the original corpus. When researchers want to use the corpus for other purposes, they can delete the original attached code and re-label it.

(2) Annotations can be extracted separately and stored separately. The attached codes used in annotation should have obviously different characteristics from the corpus itself, so that users can easily distinguish them.

(3) Corpus users should be clear about the principles of annotation and the significance of attaching codes. Since there is no standard and unified annotation at present, all annotation corpora should be equipped with detailed annotation introduction manuals for users' reference.

(4) In the instruction document of the corpus, the annotator and the method used in the annotation should be explained. At present, there are three methods to segment and label large-scale corpus, namely, manual annotation, automatic annotation or a combination of the two. It is very useful for users to understand the meaning of annotation code by explaining the annotation method.

(5) It should be shown to users that corpus annotation is not perfect, but only a tool. Whether manual annotation, automatic annotation, or a combination of the two may lead to annotation differences, because annotation is essentially an explanation of language features, and different people may have different interpretations.

(6) The generally accepted neutral mode should be adopted as far as possible. In order to facilitate the use of corpus, annotation should be based on a comprehensive use of a wide range of grammatical

ISSN 2616-7433 Vol. 5, Issue 11: 55-59, DOI: 10.25236/FSST.2023.051109

theories, not limited to a specific grammatical theory.

(7) No annotation mode can be used as the first standard. At present, there is not a generally accepted annotation model in academic circles. Therefore, the objective approach is to comprehensively investigate various existing annotation modes, learn from each other's strong points and establish a compromise annotation mode.

5. Conclusions

The correct use of English-Chinese translation bilingual corpus can greatly improve the efficiency of English translation teaching and make students really improve their translation ability. Therefore, in teaching, teachers can not only guide students to use corpus correctly, but also independently establish a memory bank and a terminology bank to further improve the learning quality. Generally speaking, in the process of constructing English-Chinese bilingual corpus, if the scale, use and scope of application are not perfect, the application and research will be limited. In the process of construction, firstly, bilingual text materials should be obtained from the official English learning website, systematically screened, corpus memory and translation terminology database should be established and improved, students should be guided to use bilingual parallel corpora more, translation rules should be summarized, and translation skills should be learned, so as to gradually form a corpus development and use environment and achieve the purpose of construction and use.

References

[1] A. B. Yu, "Establishment and Application of Chinese-English Parallel Corpora for tourism publicity: A case study of Wuxi city," English on Campus, vol. 23, no. 6, pp. 74-75+84, 2022.
[2] L. J. Liu. Application of Corpus in College English Translation Teaching[J]. Journal of Changchun University, 2022, 32(06):90-93.
[3] Z. J. Lian. A Corpus-based Study of the Construction of Cloud Platform of Translation Laboratory and its Teaching Mode [J]. Research and Exploration in Laboratory, 2022, 41(05): 230-233+262.
[4] X. F. Ge, J. Liu, "On Web Crawler Software Module Design," Journal of Heihe University, 2018, 9(10): 209-210.
[5] Wang Y, Ren Y. Data Acquisition on Network Public Opinion of Emergencies Based on Web Crawler Technology [C] //Institute of Management Science and Industrial Engineering. Proceedings of 2019 7th International Conference on Machinery, Materials and Computing Technology(ICMMCT 2019). Clausius Scientific Press, 2019: 679-683.
[6] S. Wen. Research and application of key techniques in Chinese-English parallel corpus [D]. Guangxi University for Nationalities, 2021.
[7] Rémi C, Natalia G. Parallel Sentence Alignment from Biomedical Comparable Corpora.[J]. Studies in health technology and informatics, 2020, 270(1): 362-366.
[8] H. S. Wang, S. S. Wang. Research on the translation application of corpus alignment technique [J]. Chinese Science & Technology Translators Journal, 2017, 30(04): 16-19.
[9] L. Shu, Y. L. Guo, H. P. Wang, et al. The Construction and Application of Ancient Chinese Corpus with Word Sense Annotation[J]. Journal of Chinese Information Processing, 2022, 36(05): 21-30.
[10] A. H. Dong. A Discussion to Annotation of the Corpora[J]. Journal of Beijing Institute of Graphic Communication, 2016, 24(05): 67-70.