ISSN 2616-7433 Vol. 5, Issue 18: 83-87, DOI: 10.25236/FSST.2023.051814

Construction of Parallel Corpus for Japanese Software Outsourcing Document Translation

Fangting Liu*

College of Foreign Languages, Bohai University, Jinzhou, 121013, China
liufangting1984@126.com
**Corresponding author*

Abstract: *Software service outsourcing is a kind of service trade, which is the result of the continuous refinement of social division of labor and the development of software technology. A parallel corpus is a bilingual or multilingual corpus composed of the original text and its parallel counterpart in the target language. In the process of software outsourcing to Japan, the mutual translation of software documents is one of the important tasks. This paper studies the construction of parallel corpora for software outsourcing document translation to Japan, including AntConc tool, corpus construction steps, word frequency statistics, Concordance and Concordance Plot. The research results solve the key problem of constructing parallel corpora for Japanese software outsourcing document translation. In the actual construction process, the builder only needs to choose the corresponding software or tools according to the construction scheme proposed in this topic, and it can be completed step by step. In the process of application, it is necessary to follow the actual needs of the Japanese software outsourcing document translation practice, research and talent training and other fields.*

Keywords: *Software Outsourcing for Japan; Document Translation; Parallel Corpus; AntConc; Word Frequency Statistics*

1. Introduction

Software outsourcing is a software demand activity in which an enterprise contracts all or part of the work in a software project to an enterprise that provides outsourcing services in order to focus on its core competitiveness and reduce the cost of a software project. Now business process outsourcing has become a new development trend of outsourcing services, and will become the main content of outsourcing in the next few years. With the development of social science and technology, China has gradually stepped into the information age, and the software industry has become a powerful emerging industry in China's development. In the new century, many companies in China have begun to get involved in the field of software outsourcing, and the software outsourcing industry has shown a good state of development and stepped into the growth stage of the software outsourcing industry [1]. Cities with more developed software outsourcing industries include Shanghai, Beijing, Dalian and Shenzhen. In Beijing, for example, 40% of software enterprises participate in outsourcing projects, and 60% to 70% of the turnover of the software industry comes from outsourcing. Offshore software outsourcing refers to the behavior that an enterprise subcontracts some non-core software projects to software enterprises in countries with low labor costs for development in order to achieve low-cost operation [2]. China has taken advantage of its latecomer advantage in the field of digital economy to drive the rapid growth of related industries, and China's offshore software outsourcing services have bucked the trend of growth in the downturn of total software exports.

Software documents are materials related to the development, maintenance and use of software, and are technical data and information in a form that people can read. Software project documentation is an important part of software project development, and it plays an important role in ensuring and supporting the success of project development and project maintenance. In the process of software outsourcing to Japan, the mutual translation of software documents is an important task. Sometimes the Japanese documents need to be translated into Chinese, and sometimes the Chinese documents need to be translated into Japanese. Mainly used for statistical analysis and hypothesis testing of language rules, corpora play a very important role in modern linguistic research and language education, with unique advantages such as large capacity, truthful corpus and fast and accurate retrieval [3]. A parallel corpus is a corpus composed of two or more languages, consisting of texts, paragraphs and sentences

ISSN 2616-7433 Vol. 5, Issue 18: 83-87, DOI: 10.25236/FSST.2023.051814

uniformly processed by the source language and target language. Bilingual correspondence can be divided into word level, sentence level and paragraph level. At present, the construction of parallel corpora is mostly centered on English, and there are few parallel corpora with minor languages and Chinese as the original source language and target language [4]. Building a parallel corpus of document translation for software outsourcing to Japan is an effective way to assist translators, improve translation quality, train translation talents and promote machine translation.

2. AntConc Tool

AntConc is a free corpus analysis and research tool developed by Laurence Anthony in 2002, and has corresponding versions for Windows, Mac and Linux systems [5-7]. AntCont3.2.4 software is compatible with Chinese, Japanese and Korean, and can conduct corpus statistics for UTF-8 and UTF-16 text encoding, and supports regular expressions, which is more practical and convenient than WordSmith Tools. Especially when dealing with Chinese corpus without Spaces, regular expressions can be used to add Spaces in batches. AntCont3.2.4 includes: word retrieval function, which can list all the sentences and fragments containing the searched words, so as to facilitate the analysis of the situation in which the searched words are located. The index word map function can intuitively reproduce the position and density of the searched word in the whole text. Generate word list function, sort by word frequency, clearly show the level of word frequency. Subject word function, the listed words are the words listed in the observation text and another reference text with significant differences. File viewer, click the retrieved word, you can visually observe its distribution in the source file. Word cluster function, continuous multiple words and contains a word or phrase of the text fragment, through the computer can automatically retrieve in the same form repeated 2 words or more words of meaningful continuous phrase units. The essence of the phrase collocation function is to find the collocation of the index words.

3. Corpus Construction

Although many corpora have been built, there is no corpus specifically for outsourcing document translation to Japanese software. Although there are some relevant corpora in other large corpora, searching them is time-consuming and laborious. Therefore, in order to carry out this research, it is necessary to build a corpus of document translation for software outsourcing to Japan. The construction process includes the preparation of corpus, the textualization of corpus, the import of source corpus and the de-duplication of corpus. Among them, the corpus preparation is mainly the use of previous outsourced translation documents to Japanese software. Corpus textualization converts corpus resources into the format required by AntConc tool. Source data import, the textual source data into the AntConc system. Corpus deweighting is the most important task in corpus construction, because the outsourced translation documents to Japanese software are large and there are a large number of repetitive sentences.

This study intends to adopt the Term frequency-inverse Document Frequency (TF-IDF) algorithm for similar sentence segment de-duplication in normal corpus [8,9]. Align the sentences corresponding to each other in parallel corpora, deduce the similarity degree of sentence segments, strengthen the subsequent removal speed, integrate TF-IDF technology and word theme correlation, calculate keyword weights, and delete high-weight sentence segments to achieve the purpose of removing sentence segments in parallel corpora. TF-IDF technology mainly calculates the importance of keywords in sentences, TF is the frequency of keyword occurrence in sentences, and the mathematical formula of TF describing keyword t_i is as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

In the above formula, $n_{i,j}$ represents the number of occurrences of keyword t_i in sentence segment *j*, and $\sum_k n_{k,j}$ is the total number of occurrences of all keywords in sentence segment *j*. The higher the number of keyword occurrences in the sentence paragraph, the larger the TF value. Inverse text frequency IDF, representing the proportion of the sentence segment in which the keyword is

ISSN 2616-7433 Vol. 5, Issue 18: 83-87, DOI: 10.25236/FSST.2023.051814

located in the sentence segment set, is denoted as:

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2}$$

In the above formula, $|D|$ is the total number of sentence segments in the set of t_i sentence segments, and $|\{j : t_i \in d_j\}|$ is the number of sentence segments containing keywords, and it is guaranteed that $n_{i,j}$ is not equal to zero sentence segments. The TF-IDF of keyword *i* in paragraph *j* is defined as:

$$tf - idf_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

The effectiveness of TF-IDF technology lies in that the higher the frequency of sentence segment occurrence, the lower the importance level, which has a certain bias for some sentence segments, and a certain type of key words appear a lot in the sentence segment, so more weight should be given to such words.

4. Concordance and Concordance Plot

"Concordance" is a core part of modern corpus technology. Users only need to enter the word or phrase to be searched in the search box, click the "Start" button, and all the contexts containing the word or phrase in the corpus text will immediately appear in the right window and be listed as a sort list, which is very convenient for users to understand the frequency of use of words or phrases, use methods and common collocation forms. As long as the user clicks on any item in the list on the right, the use of the word or phrase in the corresponding context in the corpus text will be clear. Users can carry out statistical analysis of the linguistic phenomena processed by searching, so as to find the linguistic laws contained therein, which has a very good auxiliary role for Japanese translation and Japanese document learning.

Move the pointer to the search term highlighted by one of the index lines, the pointer becomes a hand-shaped tool, click the search term, you can see the search term appears in the original text. The total number of index lines is displayed under "concordance hits", and at the end of the processing, "FINISHED" is displayed. If NO index row is generated, "NO HITS" is displayed, and the index row window is not updated. The search term can be set to "word (default)" or "word fragment" by the Word option above "search term", case-insensitive by "case", or full regular expression by selecting "Regex". The AntConc tool also provides advanced search terms: defining a set of search terms that can be entered one line at a time or directly into a list of search terms in a file allows users to use a set of search terms without having to type them repeatedly each time. Defines the context term and the scope of a context in which the search term must appear.

"Concordance Plot" is another retrieval aid provided by the AntConc tool. The steps for index positioning are the same as using the index tool, which provides a different view of the index row. The number of indexes in each file is located on a bar code, indicating the position of the index line in the text with the search term. You can see which files contain the target search term, and you can also determine where the search term meets the term from. The number of indexes and the file length are displayed to the right of the bar code.

5. Word Frequency Statistics

Word frequency statistics is a new method of lexical analysis and research, which is mainly used to understand word rules by counting the occurrence times and frequency of each word in a corpus with a certain capacity. Word frequency statistics are used in linguistics, information science, informatics and bibliometrics.

5.1 Principles of AntConc Word Frequency Statistics

AntConc uses the processing power and algorithm of the computer to process and analyze the input

ISSN 2616-7433 Vol. 5, Issue 18: 83-87, DOI: 10.25236/FSST.2023.051814

text and realize the function of word frequency statistics. The specific principle is as follows: First, text preprocessing. AntConc first preprocesses the input text, including removing punctuation and stopping words, in order to better analyze and count the words. Second, word segmentation. AntConc divides the processed text according to Spaces and newlines to get one word at a time. Third, word count. AntConc counts the segmented words, counting the number of times each word appears in the text. Fourth, sort and filter. AntConc sorts words according to the number of times they appear, and can choose to show only words that appear more frequently, or to exclude certain words. Fifth, visual display. AntConc displays the statistical results in the form of tables, bar charts and word clouds, which is convenient for users to analyze and understand.

5.2 Application of AntConc Word Frequency Statistics

AntConc word frequency statistics has a wide range of applications in text analysis and research. The common application scenarios are as follows: First, text mining. Through word frequency statistics, we can understand the key words or topics in a text, and help researchers quickly grasp the core content of the text. The second is linguistic research. Word frequency statistics can help researchers analyze the characteristics of a certain language, including the frequency of use of certain words and the diversity of vocabulary, so as to deeply understand the rules and characteristics of the language. Third, word processing. Word frequency statistics can help editors check for repeated words, words that are too frequent or too low in an article, thereby improving the quality and readability of the article. Fourth, education and teaching. Word frequency statistics can help teachers understand students' writing level and vocabulary, so as to carry out targeted teaching guidance and guidance. Fifth, sentiment analysis. Word frequency statistics can analyze the use of positive and negative emotional words in a text, so as to judge the emotional tendency and attitude of the text.

5.3 Advantages of AntConc Word Frequency Statistics

Compared with traditional manual statistics, AntConc word frequency statistics has the following advantages: First, automatic processing. AntConc can automatically process and analyze a large number of texts, eliminating the tedious work of manual statistics and improving efficiency. Second, high accuracy. AntConc uses computer algorithms for statistics with high accuracy, avoiding errors and subjectivity in manual statistics. Third, visual display. AntConc will display the statistical results in the form of charts, intuitive and clear, easy to analyze and understand. Fourth, strong flexibility. AntConc can be customized according to the needs of users, including selecting the scope of statistics and excluding words that do not need statistics, providing more flexibility and freedom.

5.4 AntConc's "Keyword List" Step

(1) Open AntConc, load the corpus file, click the "Keyword List" TAB, and then click "Start". If there is no reference corpus word list available, a reference corpus needs to be added.

(2) The Reference Corpus can be loaded in the "Reference Corpus", the unprocessed corpus files can be loaded, or the vocabulary of the reference corpus can be directly loaded, and the target corpus and the reference corpus can be exchanged through the exchange function.

(3) Click "Start" again, at this time, it is prompted to use the keyword list, enter the word list, click "OK".

(4) All words are sorted by word frequency by default, including statistical information such as "Rank, Freq, Keyness and Keyword". At the top of the table, you can also see the number of class characters of the corpus itself and the number of class characters of the keyword list.

(5) The sorting of words in the keyword list can be sorted according to word frequency, criticality, keyword prefix or suffix. Select "Sort by Freq/Keyness/Keyword/ Keyword End" and click "Sort". You can also select "Invert Order" to implement reverse sorting.

(6) Click on a specific word, you can jump to the search results of the KWIC mode for that word.

(7) Enter the search term in the search box to locate the word, the search also supports case-sensitive and regular expression and other advanced searches. You can also load the stop word list in advanced search, filter out unwanted words, and use "Hit Location" to turn pages up and down.

ISSN 2616-7433 Vol. 5, Issue 18: 83-87, DOI: 10.25236/FSST.2023.051814

6. Conclusions

Software documentation specifies software design details, describes software functions, and describes how to use the software. Software documentation and computer programs together form a piece of software that performs a specific function. Large software projects should have a lot of documentation related to the system. Sometimes for a small or medium-sized project, there may be thousands of lines of documentation. In the process of translation, parallel corpora can effectively assist translators to express relatively accurate translations, provide a large number of words, phrases, and even sentence pattern samples, promote the development of translation theories and breakthroughs in translation thinking, and improve translation efficiency. It needs to build a parallel corpus of software outsourcing document translation to Japan, assist translators to carry out software document translation, improve the quality of software development, and promote the high-quality development of software outsourcing industry to Japan.

Acknowledgements

This work is supported by 2023 annual social science planning fund project of Liaoning province (L23BYY014): Compilation and application of terminological dictionary for Japanese software outsourcing document translation based on corpus.

References

[1] Q. Tang. Research on the development of Chinese software outsourcing industry[J]. Intelligent City, 2016, 2(03): 26-27.
[2] Y. L. Yu, D. Q. Yang. Offshore Software Outsourcing Services Bilateral Matching Recommendation Model in the Context of Digital Economy: Based on Multi-time Goal Decision Psychological Factors [J]. Journal of Qingdao University(Natural Science Edition), 2020, 33(03): 115-122.
[3] OU-YANG Ouchun. The Application of Parallel Corpus in Translation Teaching: Taking Chinese-English Database of Chuanxilu as an Example[J]. Journal of Nanchang Normal University, 2023, 44(03): 77-80.
[4] Y. Liu, D. Y. Xiong. Construction Method of Parallel Corpus for Minority Language Machine Translation[J]. Computer Science, 2022, 49(01): 41-46.
[5] Jiang L X. Constructing an ESP Bilingual Parallel Corpus Based on AntConc: Application and Assessment[J]. Frontiers in Educational Research, 2021, 4(9): 53-58.
[6] Han X. A Study on the English Vocabulary Learning Based on the Application of Corpus Tools[J]. Lecture Notes on Language and Literature, 2023, 6(13): 109-113.
[7] Ju Q. A Corpus-based Stylistic Analysis of The Great Gatsby[J]. Academic Journal of Humanities Social Sciences, 2023, 6(16): 112-117.
[8] Drezewski A S F S C H N. Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency[J]. International Journal of Advanced Computer Science and Applications (IJACSA), 2022, 13(10): 147-151.
[9] Junsheng Z X .Extracting Keywords from Texts based on Word Frequency and Association Features[J].Procedia Computer Science,2021,187: 77-82.