Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 3712090, 9 pages
https://doi.org/10.1155/2022/3712090

Research Article

Exploration of the Problems and Solutions Based on the Translation of Computer Software into Japanese Language

Lian Hu[1] **and Jing Hu**[2]

[1]*Foreign Languages and Literatures, Wuhan University, Wuhan 430000, Hubei, China*
[2]*Foreign Languages, Wuhan University of Technology, Wuhan 430000, Hubei, China*

Correspondence should be addressed to Lian Hu; 20201010200011@whu.edu.cn

Received 5 July 2022; Revised 6 August 2022; Accepted 12 August 2022; Published 6 September 2022

Academic Editor: Baiyuan Ding

Copyright © 2022 Lian Hu and Jing Hu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

At present, the research on machine translation mainly focuses on English-Chinese translation, while the research on Japanese college students using Japanese Chinese machine translation software is relatively few. In order to solve the above problems in Chinese Japanese bilingual translation, this paper proposes a phrase translation method based on sequence intersection. This method regards sentences as word sequences and aligns the sequence intersection of all source sentences corresponding to the target sentence in the corpus with Chinese and Japanese sentences containing the phrases to be translated. By fully mining the information of sentence alignment bilingual corpus without word alignment resources, we can obtain high-quality phrase translation, syntactic analysis, and dictionary. Then, we focus on the automatic construction of sentence level aligned bilingual corpus and explore the automatic sentence alignment technology of Chinese and Japanese bilinguals. A ten-year alignment model based on combination cues and core extended square matching is proposed. The preprocessing of the computer corpus and the basic construction of the corpus are completed. This paper also puts forward corresponding countermeasures and approaches to the problems encountered in the construction of computer translation.

1. Introduction

Computer software translation (machine translation) is an important area of natural language understanding. In the early 1930s, the French scientist Altruinhad the idea of "machine translation" [1]. China began to study machine translation in 1956, and with the rapid development of the Internet and the expansion and deepening of cross-cultural communication, machine translation is gradually becoming an important means for people to overcome the language barriers they face in accessing information, and the increasing demand for translation makes translation software enter a new period of development.

With the expanding and deepening economic and cultural exchanges between China and Japan, the demand for rapid and accurate access to Japanese information in China is increasingly urgent. In machine translation, Chinese-Japanese translation has attracted much attention. Although the number of Japanese learners in China is increasing, and China has now surpassed Korea as the country with the largest number of Japanese learners, it is still far from meeting this demand [2]. The emergence of Japanese-to-Japanese machine translation software provides an efficient means to break through the Chinese-Japanese information barrier. The translation quality of today's Chinese-Chinese translation software has come a long way compared with that of the previous, but it is undeniable that there is still a huge gap between the translation quality of Chinese-Japanese translation software and that of professional translators. Especially in the field of literary translation, machine translation is still difficult to be involved due to the lack of human feelings, the lack of delicate human deduction efforts, and the lack of human "encyclopedic knowledge" and linguistic and cultural knowledge. In addition, in the field of legal documents and contracts, which requires very precise language where language errors can have serious

consequences, machine translation should be used with caution. The advantage of machine translation lies in "quantity," in its unreachable human speed, and the advantage of human translators lies in "quality." In order to translate quickly and well, we should combine the two organically and combine their outstanding advantages, such as arranging human translators to carefully review and reprocess the preliminary processed text translated by machines rapidly, so as to achieve complementary advantages [3]. However, the current research has not been able to find a solution for the Japanese and Chinese machine translation. However, the current research is relatively rare in terms of analysis and countermeasure research on the shortcomings and solutions of Japanese-Chinese machine translation software in translation.

The innovative contribution of this paper is to propose a phrase translation method based on sequence intersection. This method does not need auxiliary word alignment, syntactic analysis, and dictionary to obtain candidate translations. Sentences containing phrases to be translated can be intersected in a sentence level aligned bilingual corpus. Then, the translation of phrases is obtained after postprocessing. The phrase translation acquisition method based on sequence intersection gets rid of the dependence on word alignment, syntax analysis, and dictionary. It makes full use of the information of bilingual sentence alignment corpus and has a high accuracy. This will help build a Chinese Japanese machine translation corpus and solve the current problem of computer software corpus for Chinese Japanese translation.

2. State of the Art

Today's society is in the information age, with the rapid development of the Internet, and there is an urgent need to remove the textual barriers between people of different nationalities through machine translation. But natural language translation is one of the advanced levels of human intelligence activities, and artificial intelligence research has not yet reached the level of fully understanding natural language, so machine translation research is an important element of computational linguistics research with a significantly socioeconomic value [4]. According to the different knowledge representation and processing methods, there are two main translation methods: rule-based machine translation method and corpus-based machine translation method [5].

2.1. Rule-Based Machine Translation Methods. Until the 1990s, rule-based approaches dominated machine translation. Rule-based machine translation started with Chomsky's formal linguistics and the rise of artificial intelligence. The traditional rule-based machine translation process includes steps such as lexical analysis, syntactic analysis, semantic analysis, pragmatic analysis, intermediate language generation, and target language generation [6]. The rule-based machine translation system is to analyze, judge, and take the lexical, syntactic, semantic, and syntactic aspects of language utterances and then rearrange and combine them to generate equivalent target languages. Rule-based machine translation system can be divided into three kinds from the architecture: direct translation method, conversion generation method, and intermediate language translation method [7]. In the direct translation method, some electronic dictionary resources are mainly used, and these resources are mainly constructed by experts manually, and the translation method is also mainly based on dictionary-to-translation. As can be imagined, the quality of such translation results is not high, and the readability is poor. There are three basic steps in the conversion-based approach: analysis, conversion, and generation [8]. Analysis means transforming the source language into a predefined intrinsic abstract representation through word form processing, word division and lexical annotation, syntactic analysis, semantic analysis, etc.; conversion means transforming this intrinsic structure of the source language lexically and structurally into the corresponding target language or the abstract intrinsic structure of the target language; and generation means observing the necessary syntactic, semantic, and pragmatic constraints to produce the target language from its intrinsic structure out of the target language. The intermediate language-based approach is to analyze the source language to produce a representation that becomes an intermediate language and then to generate the target language directly from this intermediate language representation. An intermediate language is a systematic computer representation of a natural language, which attempts to create an artificial language that is independent of, and at the same time capable of representing, a variety of natural languages.

The rule-based machine translation approach can express the linguist's knowledge more intuitively and facilitate the handling of complex structures and deep understanding, which can effectively solve the long-distance dependency problem [9]. However, the rule-based approach also has its inherent drawbacks. Firstly, because the rules are handwritten by experts, they are inevitably highly subjective, and there is often a gap when these highly subjective rules are used to describe objectively existing linguistic phenomena. Secondly, the coverage of the rules is poor, as it is often difficult to summarize the rules comprehensively because these limited rules written manually are used to describe all the phenomena of a language. Although this problem can be remedied by continuously increasing the number of rules, after the rule base reaches a certain size, researchers find that the phenomenon of rule conflicts tends to be serious, making it difficult to effectively expand the existing rule base. Moreover, it is difficult to deal with real texts with rule bases constructed manually by experts, so the resulting translation systems tend to achieve better results only in some small areas [10]. Moreover, because the rule base must be constructed manually by experts with profound linguistic knowledge, it results in a high development cost of the system. The high cost of manual rule writing, the difficulty of guaranteeing rule consistency, and the difficulty of building and maintaining the relevant knowledge base make the method unable to adapt to the needs of large-scale data development and eventually escape the fate of being replaced.

2.2. A Corpus-Based Approach to Machine Translation. By the mid-to-late 1980s, the corpus-based machine translation approach gained significant development and gradually took a dominant position. This approach is based on a large-scale collection of bilingual corpora that are translated into each other. The method is further divided into two types: the statistical-based translation method (SBMT) and the instance-based translation method (EBMT) [11]. The statistical-based translation method (SBMT) was first proposed by Weaver in 1949, but statistical machine translation did not form a systematic theoretical framework in the following decades. Moreover, due to the limitations of computer computing power and corpus size at that time, researchers were not in a position to experiment with such an over-the-top theory.

The value of the IBM model is that, on the one hand, it describes machine translation in terms of a formal mathematical model for the first time and gives an effective method for estimating the model parameters; on the other hand, it provides a word-based alignment model that can automatically obtain word alignment information through training [12]. The IBM model, as the earliest proposed statistical machine translation model, has been influencing the later research on statistical machine translation. The straight IBM model as an early statistical machine translation model brought a new vision to the field of machine translation when it was first proposed, but researchers soon discovered its shortcomings. The most representative one is the limitation of the IBM model on word alignment, which does not allow one-to-many alignment cases from the source to the target language. This also limits the translation capabilities of the IBM model. In traditional machine translation systems, translation knowledge is expressed in the form of rules, which are written manually by linguists. This method requires a lot of money and manpower to develop dictionaries and rule systems. From the perspective of research, the experiment relies too much on the knowledge and experience of language rule developers, the research cycle is too long, and there is a lack of comparability between different research works. From the experimental point of view, when dealing with large-scale real data, the effect is always very unsatisfactory. With the development of machine translation research, people gradually realize that this manual method of obtaining translation rules has become a bottleneck restricting the development of machine translation research. Nowadays, phrase-based statistical machine translation has become a widely used approach in statistical machine translation, and phrase-based statistical machine translation systems have achieved good results in many reviews [13]. However, phrase-based statistical machine translation suffers from poor generalization ability, inability to represent translations of discontinuous phrase collocations, and inability to perform long-range discourse order adjustment. The best way to solve these problems is to introduce syntactic structure and build a statistical translation model based on syntactic structure, that is, syntax-based statistical translation model. And syntax-based machine translation is the current research hotspot in the field of statistical machine translation. There are numerous syntax-based statistical machine translation models, which can be broadly divided into formal syntax-based models and linguistic syntax-based models. The method used in the formal syntax-based model is the syntactic structure obtained from the corpus by automatic learning methods. The syntax used in linguistic syntax-based models is what linguists call syntactic structure and usually requires either learning the syntactic analysis of the sentence patterns in question from a manually constructed corpus or using artificially constructed linguistic rules for syntactic analysis [14].

The advantage of statistical machine translation over other machine translation methods is that it does not require the support of linguistic knowledge. The entire translation process is simulated by a mathematical model and can be done automatically by a computer. In addition, statistical machine translation requires only a certain size of bilingual aligned corpus as training data and does not rely on expensive human word-aligned corpus [15].

The instance-based translation approach obtains the target translation mainly by finding the most similar translation instances in a bilingual corpus. In instance-based machine translation systems, translation knowledge is represented as a lexicon of instances and sense classes and is easy for addition or removal, the system is simple and easy to be maintained and has the potential to produce high-quality translations if a larger library of translation instances is utilized and accurately compared, and it avoids the difficulties of deep linguistic analysis that those traditional rule-based machine translation methods must perform, which is very strategic in translation attractive [16]. In order for us to be able to accurately find the corresponding target language example sentence from the source language example sentence in the instance base, the concrete implementation of an instance-based machine translation system requires the ability to perform automatic bilingual alignment correctly, and not only at the sentence level, but also at the lexical level and even at the phrase level.

2.3. Analysis of Machine Translation Problems in Chinese and Japanese Languages. Japanese-Chinese/Chinese-Japanese machine translation is more difficult compared with Japanese-English/English-Japanese machine translation. The reason is that Chinese is an isolated language with a different grammatical structure, no active forms and tenses, and sentences cannot be divided by words, but by a string of consecutive Chinese characters, which are completely different from European and American languages and Japanese. Therefore, it is difficult to segment Chinese sentences by words, and the analysis of Chinese syntactic structure is not simple [17].

In order to continue the development of Japanese-Chinese and Chinese-Japanese machine translation, there is an urgent need to integrate machine translation research results, various accurate dictionaries, aligned parallel text data, etc. from China and Japan to build a practical Chinese-Japanese and Japanese-Chinese machine translation system. Such a high-quality and robust machine translation system

will not only enhance the development of machine translation research itself, but also apply it to cross-lingual information services between libraries, or to practical scenarios in economic and cultural exchanges, which will be a great contribution to all aspects of information exchange and resource sharing between China and Japan [18]. The translation systems developed so far are mostly computer software translation systems. Most of the currently developed computer software translation systems are based on the framework of example-based machine translation, which absorbs the complex variations of linguistic representations through the effective use of dependency structure analysis to produce high-quality translations. Future work will focus on improving the accuracy of the dependency structure analysis for Chinese and expanding the instance dictionary to several times the number of dictionaries in the existing system with a view to achieving a diversity of representations in response to language. To this end, a large collection of aligned parallel texts is required.

3. Methodology

At present, corpus-based machine translation is still the dominant approach to machine translation, and this translation method based on large-scale real text processing is still the general feature of current machine translation. Phrase translation acquisition, as the main method for constructing the unmatched part of the translation unit, is one of the indispensable core aspects of machine translation [1]. Establishing sentence-level correspondence for bilingual texts means determining which sentence or sentences in the source language text and which sentence or sentences in the target language are mutually translated. The purpose of sentence alignment is to identify the sequence of sentence beads in a bilingual text that are composed of sentences that are mutually translatable.

3.1. Phrase Translation Acquisition Method Based on Sequence Intersection. The phrase translation acquisition method based on sequence intersection consists of a basic model, a high-frequency interference word restriction module, and a support degree restriction module. The basic model extracts high-quality phrase translation pair candidates from the sentence-level aligned bilingual corpus and ranks them; the high-frequency word restriction module solves the high-frequency word interference problem in the output results of the translations; the support restriction module controls the number of output results [19].

The corpus used in the sequence intersection-based phrase translation acquisition method is the sentence-level aligned bilingual corpus BC, which contains several Chinese-Japanese aligned sentence pairs. A sentence pair S is denoted as

$$S = CS{\leftrightarrow}JS, \tag{1}$$

where CS and JS are mutually translated Chinese sentences and Japanese sentences. In this method, the sentences are represented in the form of word sequences.

$$\begin{aligned} CS &= < c_1, c_2, \ldots, c_m >, \\ JS &= < j_1, j_2, \ldots, j_n >. \end{aligned} \tag{2}$$

Thus, the sentence pair S can be expressed in the form of a word sequence.

$$S = < c_1, c_2, \ldots, c_m > \leftrightarrow < j_1, j_2, \ldots, j_n >. \tag{3}$$

Let P be the Chinese phrase to be translated, expressed in the form of a sequence of words.

$$P = < P_1, P_2, \ldots, P_n >. \tag{4}$$

Let the double statement pair $S_k, S_h \in BC$,

$$\begin{aligned} S_k &= CS_k{\leftrightarrow}JS_k = < c_{k+1}, c_{k+2}, \ldots, c_{k+m} > \leftrightarrow < j_{k+1}, j_{k+2}, \ldots, j_{k+m} >, \\ S_h &= CS_h{\leftrightarrow}JS_h = < c_{h+1}, c_{h+2}, \ldots, c_{h+m} > \leftrightarrow < j_{h+1}, j_{h+2}, \ldots, j_{h+m} >. \end{aligned} \tag{5}$$

S_k intersection with S_h is defined as

$$S_k \cap S_h = CS_h \cap CS_k {\leftrightarrow} JS_h \cap JS_k, \tag{6}$$

where $CS_h \cap CS_k$ is defined as

$$\begin{aligned} &CS_h \cap CS_k \\ &= \arg\max_{\langle c_{h+h1}, c_{h+h2}, \ldots, c_{h+hq} \rangle} \left| \langle c_{h+h1}, c_{h+h2}, \ldots, c_{h+hq} \rangle \right|, \end{aligned} \tag{7}$$

$$0 \le h_1 < h_2 < \ldots h_q \le m_h, \tag{8}$$

$$0 \le k_1 < k_2 < \cdots k_r \le m_k. \tag{9}$$

Equation (7) indicates that the result of $CS_h \cap CS_k$ is a new word sequence, and each word in this sequence corresponds to each word in CS_h, CS_k , and the subscripts $h1, h2, \ldots, hq$ and $k1, k2, \ldots, kr$ of the two sequences should fall within the subscripts of CS_h, CS_k and be monotonically increasing, respectively; that is, they should satisfy (8) and (9).

If the intersection of S_k and S_h is

$$S_k \cap S_h = P{\leftrightarrow}T = P{\leftrightarrow} < j_{g1}, j_{g2}, \ldots, j_{gn} >, \tag{10}$$

P is the Chinese phrase to be translated and T is the intersection of the Japanese parts of S_k and S_h. Then, say that S_k and S_h support $P{\leftrightarrow}T$, and call T a candidate translation for P. If there are x sentence pairs in the corpus supporting $P{\leftrightarrow}T$, then the support of T as a candidate translation of i is said to be x, denoted as

$$SV(P{\leftrightarrow}T) = x. \tag{11}$$

The candidate translation with the highest support was selected as the translation result for P.

$$\text{Translation}(P) = \arg\max SV(P{\leftrightarrow}T). \tag{12}$$

In general, the translation of phrases results in a certain tendency of continuity in the translated sentences. And this tendency is not reflected in the basic model. If $g1, g2, \ldots, gn$ is continuous, it is a strong candidate translation of P; otherwise, it is an if candidate translation. In the continuity-

constrained model, the strong candidate translation with the greatest support is chosen as the translation result of P; if there is no strong candidate translation, the weak candidate translation with the greatest support is chosen as the translation result of P.

3.2. Length-Based Sentence Alignment Methods. By alignment, we mean the creation of a mapping of intertranslational fragments or units between two languages in a parallel corpus. In simple terms, the sentence alignment problem is the process of corresponding a set of sentences in the source language to a set of sentences in the target language in terms of sentence content. As a special kind of corpus, parallel corpus is important for research on corpus-based machine translation, human-machine interactive translation, machine translation evaluation tools, cross-lingual information retrieval, bilingual phrase dictionary compilation, and word sense disambiguation [20].

If one wants to obtain a larger bilingual knowledge base, one must first establish the sentence-level pairwise translation relations of the obtained bilingual texts. And the correspondence relations between bilingual sentences include complex forms of one-to-many and many-to-many, in addition to a large number of one-to-one cases, and are thus quite technically challenging. BMT uses the mechanism of analogy for natural language understanding, which does not require understanding of the source language but requires keeping a large library of instances in which a large number of bilingual contrastive sentences or phrases are kept. When a sentence needs to be translated, the system goes to the instance library to find one or more source language instances that are similar or partially similar to it, identifies its corresponding target language instances, represents the sentence as some combination or transformation of these source language instances, and then applies the same combination or transformation to the target language corresponding to these instances to obtain a target language translation of the sentence [21].

This paper presents a bilingual sentence alignment method using sentence length and position information. This is a sentence alignment method for bilingual texts, wherein a plurality of alignment anchor points is calibrated in the bilingual text before automatic alignment. The alignment anchor divides the bilingual text into several alignment intervals, and automatic alignment is carried out within the several alignment intervals, respectively. The sentence alignment method of bilingual text according to claim 1: the alignment anchor points are uniformly distributed and calibrated in the bilingual text. The sentence alignment method of bilingual text according to claim 1 or 2: after the automatic alignment is performed, the sentence alignment results in the alignment interval are checked, and the alignment anchor points incorrectly calibrated in the automatic alignment process are modified and calibrated. In length-based alignment methods, some use the number of words in a sentence as a measure of sentence length units, while others use the number of characters in a sentence as a measure of sentence length. Such alignment algorithms require no linguistic knowledge and utilize a very simple statistical model based mainly on the fact that long sentences in one language are still longer when translated into the other language; conversely, short sentences are still shorter when translated into the other language; that is, the two mutually translated sentences are considered to be highly correlated in length. Another premise of this alignment algorithm is that the order of the mutually translated sentences does not change drastically in their respective texts. The length-based approach treats sentence alignment as a function of sentence length, does not require additional lexical information, and is more efficient, but is prone to error spreading.

The number of characters C corresponding to each character in language $L1$ in language $L2$ is a random variable and that random variable is normally distributed $N(c, s2)$, defined as

$$\delta = \frac{(l_2 - l_1 c)}{\sqrt{l_1 s^2}}. \tag{13}$$

For the probability $p(\text{match}|\delta)$, this is transformed into $p(\delta|\text{match})p(\text{match})$ using the Bayesian formula, where P(match) is a constant and can be statistically derived from the tagged corpus, and $p(\delta|\text{match})$ can be estimated by the following equation:

$$p(\delta|\text{match}) = 2(1 - p(\delta)), \tag{14}$$

of which

$$p(\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} \mathrm{d}z. \tag{15}$$

Based on the above distance metric, the distance values are calculated for various alignment cases, and then, the dynamic programming algorithm is used to determine the sentence alignment of the two texts by calculating the minimum distance between the two texts.

3.3. Implementation of Segmentation and Alignment Tools for Chinese and Japanese Language Corpora. Bilingual corpus is a kind of corpus containing information of mutual translation between two languages. It can provide rich matching information between two languages and has important applications in the fields of translation knowledge acquisition, bilingual dictionary building, instance-based machine translation, word sense disambiguation, etc. In computer language compilation, alignment mostly adopts a stepwise sequential alignment method based on the length of the original translation. It is assumed that the two languages correspond in a sequential and equal proportion, and for each alignment of a language unit referring to a chapter, paragraph, or sentence, the remaining language units are then aligned according to the new proportion. In the collection of corpora, the practical needs of language research and natural language processing research and application are considered. While paying attention to the scale and the quality of the original text and the translation, the balance of various genres and chronological data should also be fully

considered. In order to meet the needs of knowledge extraction in natural language research, the corpus collected in the translation corpus is processed in three aspects: original translation alignment, part of speech tagging, and syntactic tagging.

This software tool has two main functional modules: the segmentation module and the alignment module. The segmentation module is responsible for segmenting the Chinese and Japanese English source files in sentence units, where the Chinese and Japanese corpora are separated by the Chinese period (".") as the separator. After segmentation, each natural sentence is a natural segment, and the segmentation result is saved as a corresponding document with the following file name: original file name-after segmentation. extension. The alignment module realizes the function of aligning the translation of Chinese and Japanese, Chinese, and English files by sentence and saves each group of alignment results to an excel file with the file name: original file name-alignment. During the processing of the two modules, the similarity between the generated file name and the original file name is maintained to improve the user's experience. The processing flow of each module is shown in Figures 1 and 2.

4. Result Analysis and Discussion

In order to verify the effectiveness of the basic model of phrase translation acquisition based on sequence intersection proposed in this paper, and the improvement of the basic model by the high-frequency interfering word restriction module and support restriction module, the sentence-aligned bilingual corpus *BC* used in this experiment is a 1000-sentence Chinese-Japanese bilingual sentence-aligned corpus in the sports domain. We randomly selected 40,000 Chinese phrases from *BC* for testing. Also, to test the effectiveness of our designed parallel corpus-based Chinese and Japanese language computer software translation, we experimentally compared the machine translation results with the human translation results.

In order to ensure the accuracy of the experiment, the order of the various categories of corpus selected from the corpus was disordered and not sorted according to categories, nor was the content to be translated informed. The test data used in the experiment was provided by Fuji Xerox, and we randomly selected eight chapter-level alignments in different domains and manually marked the standard sentence alignment answers. A total of 512 alignments were included in the standard answers of the test set.

4.1. Sample Validity Analysis. In order to verify whether the hypothesis of this experiment is valid, we need to test the sample. First of all, the samples of this experiment are two independent samples of machine translated translations and human translated translations, and the sample size is less than 30; if it meets the normal distribution, then it meets the *t*-test criteria; then, first of all, we judge whether the samples meet the normal distribution. We use the software SPSS to do the normal analysis for the human translation sample and the machine translation sample, respectively, and the results are shown in Figures 3 and 4.

According to the normality test criteria, it is known that the normal distribution is met when the significance value in Kolmogorov–Smirnov is greater than 0.05, and from the experimental data, we find that the significance value of both human translation and machine translation is 0.200, and 0.200 is greater than 0.05, which means that the normal distribution is met, and from the Q-Q plot, we can see that most of the experimental data are near the straight line, so the human translation and machine translation samples both conform to the normal distribution characteristics. Therefore, this experiment meets the requirements of independent sample *t*-test, and independent sample *t*-test can be conducted.

4.2. Comparison and Analysis of Experimental Results. To measure the translation results of translation software for different sentence types, we used Gale's sentence alignment system, which is very influential in the field of sentence alignment, as the Baseline system. Eight types of chapter-level alignment subaccounts were also selected from the corpus, and the results of the length comparison of different sentences in the article for the Chinese and Japanese languages are given in Figure 5.

In order to measure the translation results of the translation software for different sentence types, three sentences were selected for each sentence in this experiment, and the average scores of each of the three sentences were used to represent each score of such sentences (e.g., if the fidelity score of all three sentences is 5, then the score of such sentences in terms of fidelity is $(5 + 5 + 5)/3 = 5$, collectively referred to as the fidelity score).

From Figure 6, it can be seen that the mean scores of fidelity, natural mean, and total mean scores for human translation are 4.57, 4.14, and 4.40, respectively, and the mean scores of each item for machine translation are 3.33, 3.00, and 3.20, respectively. It can be seen from the figure that the computer software translation designed in this paper is close to the level of professionals in terms of scores.

A comparison of the number of correct alignments between these two systems on multiple alignment types is given in Figure 7, from which it can be seen that the advantage of the sentence alignment system is more obvious on multiple alignment types. This further validates that the use of the combined cue-based similarity calculation method makes full use of the connection between the two and explores the connection between Chinese and Japanese bilinguals more comprehensively and fully than the length-based method alone. Some rely on syntactic analysis or word alignment technology, which requires high resources. In this paper, a phrase translation acquisition method based on sequence intersection is proposed. Without the help of word alignment, syntactic analysis, and dictionary, this method can find the intersection of sentence pairs containing phrases to be translated in the sentence level aligned bilingual corpus. The candidate translation is obtained, and then the translated translation of the phrase is obtained through

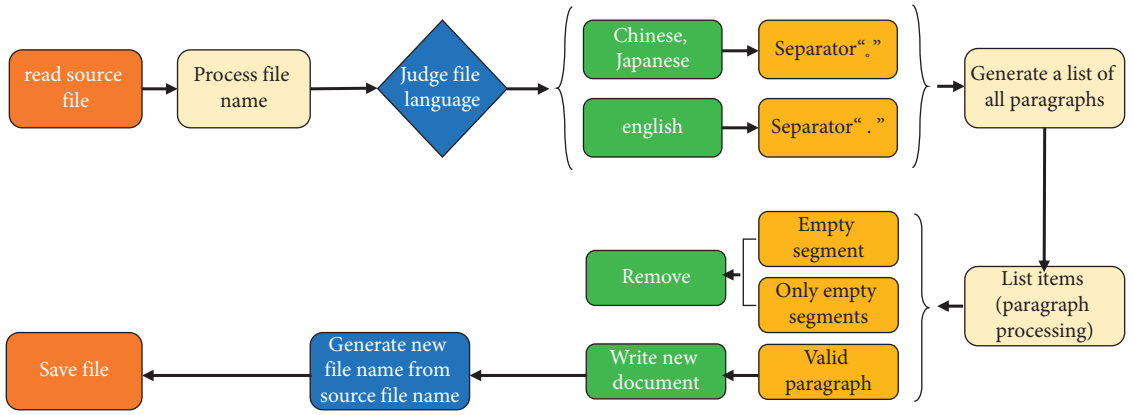

Figure 1: Flow chart of segmentation function module.

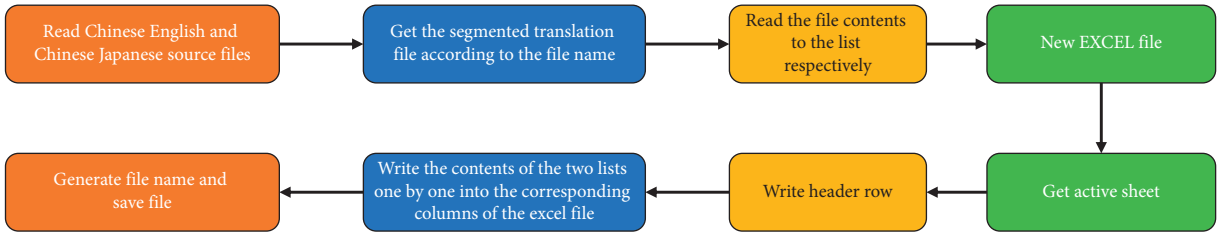

Figure 2: Flow chart of the alignment function module.

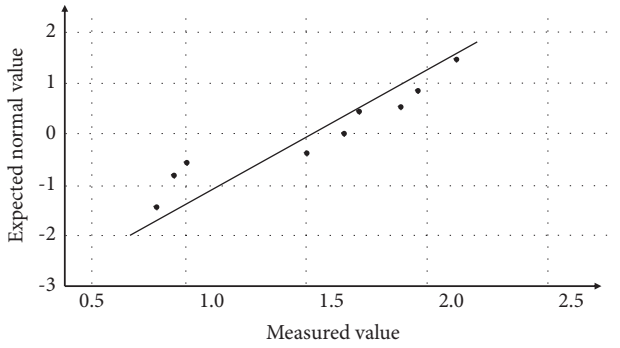

Figure 3: Analysis of human translation normality test.

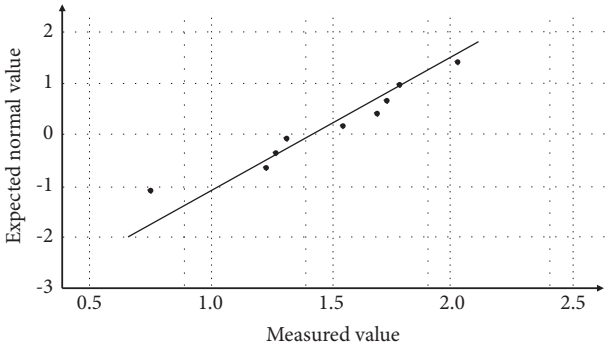

Figure 4: Analysis of machine translation normality test.

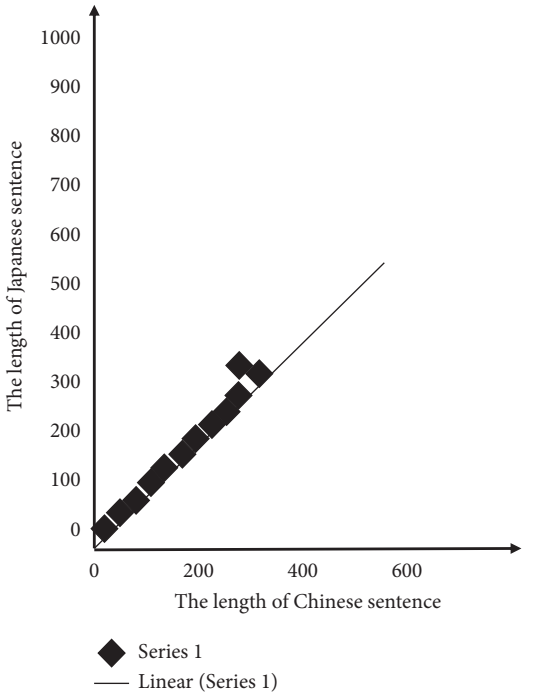

Figure 5: Ratio of sentence length between China and Japan.

postprocessing. The phrase translation acquisition method based on sequence intersection gets rid of the dependence on word alignment, syntax analysis, and dictionary. It can be used as a module of multistrategy phrase translation acquisition.

The translation quality of today's Chinese-Chinese translation software has made great progress compared with that before, but it is undeniable that there is still a huge gap between the translation quality of Chinese-Japanese

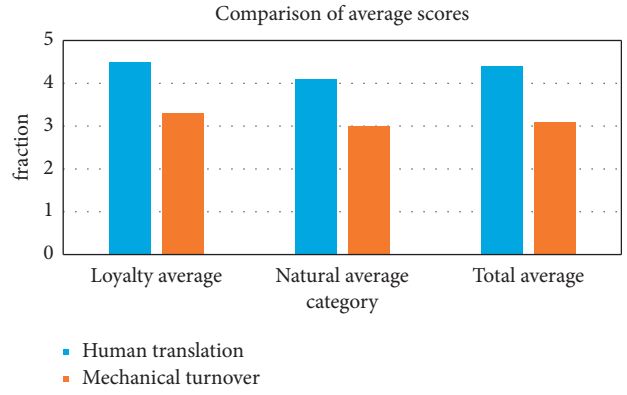

Figure 6: Score results of human translation vs. machine translation.

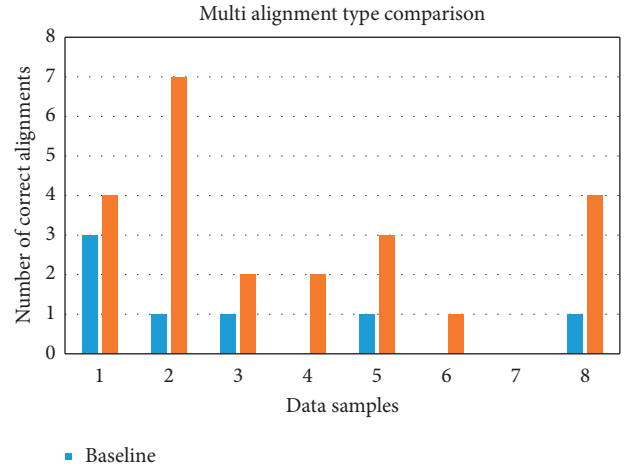

Figure 7: Results of sentence alignment on different test data.

translation software and that of professional translators. By building a large-capacity parallel corpus, it helps improve the quality and effectiveness of computer software translation.

5. Conclusion

This paper explores the automatic sentence alignment technology of Chinese and Japanese bilinguals and proposes a ten-year alignment model based on combination cues and core extended square matching. In this model, the similarity between bilingual sentences is calculated by using dictionary, word form, length, and special character combination clues to construct a ten-year aligned similarity matrix. This method does not need auxiliary word alignment, syntactic analysis, and dictionary to obtain candidate translations. Sentences containing phrases to be translated can intersect in a sentence level aligned bilingual corpus. And then, the phrase translation is obtained through postprocessing. The phrase translation acquisition method based on sequence intersection gets rid of the dependence on word alignment, grammar analysis, and dictionary. It is further verified that the combination similarity calculation method based on clues makes full use of the relationship between the two and explores the relationship between Chinese and Japanese bilinguals more comprehensively and comprehensively than the length-based method alone. However, the research has certain limitations. The research also needs to collect a large number of aligned parallel text comparisons, and future work needs to focus on improving the accuracy of Chinese dependency structure analysis. And expand the example dictionary to several times the number of dictionaries in the existing system to realize the representation of language diversity.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

References

[1] J. Zhang and T. Matsumoto, "Corpus augmentation for neural machine translation with Chinese-Japanese parallel corpora," *Applied Sciences*, vol. 9, no. 10, p. 2036, 2019.

[2] T. T. Du, "Teaching and Learning Literature in the English language curriculum in Vietnamese university education: problems and solutions," *Psychology Research*, vol. 12, no. 5, p. 10, 2022.

[3] N. Tangtorrith and N. Pongpairoj, "Systematicity of L2 interlanguage of stress assignment in English compound nouns and phrasal verbs by L1 Thai learners," *LEARN Journal: Language Education and Acquisition Research Network*, vol. 15, no. 1, pp. 33–63, 2022.

[4] Y. Hao, H. Kaiyu, W. Yu, and H. Degen, "Lexicon-augmented cross-domain Chinese word segmentation with graph convolutional network," *Chinese Journal of Electronics*, vol. 31, no. 5, pp. 1–10, 2022.

[5] H. Mezawa, S. Aoki, S. F. Nakayama et al., "Psychometric profile of the ages and stages questionnaires, Japanese translation," *Pediatrics International*, vol. 61, no. 11, pp. 1086–1095, 2019.

[6] M. Koponen, L. Salmi, and M. Nikulin, "A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output," *Machine Translation*, vol. 33, no. 1-2, pp. 61–90, 2019.

[7] J. X. Huang, K. S. Lee, and Y. K. Kim, "Hybrid translation with classification: revisiting rule-based and neural machine translation," *Electronics*, vol. 9, no. 2, p. 201, 2020.

[8] B. Lu and E. Niyomsilp, "Survey on attitude towards new management mode of linguistic landscape programs in Chinese university language resources and its decision tree analysis," *English Linguistics Research*, vol. 10, no. 4, pp. 1–21, 2021.

[9] R. Kaladevi, A. Revathi, and A. Manju, "Analyzing the evolution of modern Tamil script for natural language processing," *ECS Transactions*, vol. 107, no. 1, pp. 5219–5226, 2022.

[10] S. I. Shozamonov, S. A. Nazarova, and B. B. Djuraev, "Problems of development of the Uzbek language in current society," *Open Journal of Modern Linguistics*, vol. 11, no. 04, pp. 613–620, 2021.

[11] H. Sun, R. Wang, M. Utiyama et al., "Rui wang, masao utiyama, benjamin marie, kehai chen, eiichiro sumita, tiejun zhao. Unsupervised neural machine translation for similar and distant language pairs: an empirical study," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1–17, 2021.

[12] S. Sun, "The influence of the pluralism of Chinese language and literature on the tradition of literary criticism," *Journal of Contemporary Educational Research*, vol. 4, no. 8, pp. 91–94, 2020.

[13] Y. Ozaki, T. Goto, M. Kobayashi, and G Kutsuzawa, "Reliability and validity of the Japanese translation of brief self-control scale (BSCS-J)," *Japanese Journal of Psychology*, vol. 87, no. 2, pp. 144–154, 2016.

[14] X. Xu, "An empirical study based on the teaching quality and related issues of Japanese education in colleges and universities," *Journal of Contemporary Educational Research*, vol. 4, no. 1, p. 5, 2020.

[15] E. Sato, K. Matsuda, and B. J. Carducci, "A factor analytical investigation of the Japanese translation of the Cheek-Buss Shyness Scale in support of the three-component model of shyness," *Personality and Individual Differences*, vol. 124, pp. 160–167, 2018.

[16] P. Razmdideh, A. A. Ahangar, S. M. Sabbagh-Jafari, and G. Haffari, "An efficient method to add chunker rules in Persian to English rule-based apertium machine translation system," *Translation Studies Quarterly*, vol. 17, no. 65, pp. 54–73, 2019.

[17] W. Luo, "Analyzing the problems of vocabulary in Japanese-Chinese neural network machine translation," *Computer Science and Application*, vol. 10, no. 3, pp. 387–397, 2020.

[18] V. Liubinienė, D. Lisaitė, and J. Motiejūnienė, "A snapshot of children's attitudes toward machine translation," *Information*, vol. 13, no. 7, p. 317, 2022.

[19] M. Oe, Y. Kobayashi, T. Ishida et al., "Screening for psychotrauma related symptoms: Japanese translation and pilot testing of the Global Psychotrauma Screen," *European Journal of Psychotraumatology*, vol. 11, no. 1, Article ID 1810893, 2020.

[20] X. Wang, C. Chen, and Z. Xing, "Domain-specific machine translation with recurrent neural network for software localization," *Empirical Software Engineering*, vol. 24, no. 6, pp. 3514–3545, 2019.

[21] H. Duan, C. Zhu, and Y. Chen, "A study of the relationship between Japanese viewpoint theory and the sentences of giving and receiving, passive sentences and moving sentences based on computer software translation," *Journal of Physics: conference Series. IOP Publishing*, vol. 1744, Article ID 032085, 2021.