Enlightenment of corpus-assisted medical translation based on computer technology

Chen Chen¹ and Zikai Guo¹,*

¹Center for Medical Language and Culture Study, Xi’an Medical University, Shaanxi Province

*Corresponding author’s e-mail: guozikai@xmu.edu.cn
*Corresponding author’s ORCID: https://orcid.org/0000-0002-8592-3503

Abstract. With the advancement of modern science and technology, more and more computers or corpus aids used by computers or other equipment are used in translation more and more, especially in some professional fields. Due to a large number of professional words and complex topic sentence propositions, corpus-assisted translation has played an increasingly important role in our scientific translation at this stage. Furthermore, with the continuous expansion of our international medical program exchanges, the translation of medical English plays a significant role in developing our country’s medical and health services and promoting other medical information. Traditional English translation cannot meet the growing demand for translation, and computer-assisted translation is a trend in translation. This paper studies the corpus-assisted medical translation based on computer technology. Firstly, it uses the literature research method to explain the characteristics of medical translation and the advantages of corpus-assisted translation, and then gives relevant suggestions for auxiliary medical translation. Finally, experiments are carried out on computer-aided corpus-assisted medical translation, and changes in the efficiency and quality of translation after medical-assisted translation are added. The experiment results show that the efficiency of the medical-assisted translation is not improved, but the quality of the translation is better. The accuracy percentage of the experimental group's translation term is 87% higher than that of the control group 65%.

1. Introductions

Medical translation is highly professional. With the approaching of medical exchanges between countries, new medical treatment programs and medical machines with artificial intelligence are added, and medicines developed in research and development continue to appear and be updated [1-2]. Chinese medical staff not only need to translate these emerging treatments to learn their treatment plans [3-4], but also need to translate China’s newly discovered technologies into English and promote China’s advanced medical technologies in other countries. [5-6], which greatly increased the demand for medical translation. In recent years, corpus-assisted medical translation has developed rapidly [7-8]. In the translation of professional documents and major projects that have been introduced, CAT software has great advantages over manual translation [9-10].

There are many related types of research on computer technology corpus-assisted medical translation. First of all, researchers have raised two questions about translation: Can machine-assisted translation tools ensure the consistency of terminology translation in scientific English translation? Secondly, when the matching rate is greater than how much, the quality of machine-aided translation tools is better than machine translation [11]. For medical translation, the analysis method and the structural
model proposed by related researchers are used as the theoretical framework for comparative genre analysis. The common language steps and language features of clinical medical research papers in English and Chinese are summarized, and the clinical medical research papers in English and Chinese are found[12].

This paper studies the corpus-assisted medical translation based on computer technology. Firstly, it uses the literature research method to explain the characteristics of medical translation and the advantages of corpus-assisted translation, and then gives relevant suggestions for auxiliary medical translation. Finally, experiments are carried out on computer-aided corpus-assisted medical translation, and changes in the efficiency and quality of translation after medical-assisted translation are added.

2. Corpus-Assisted Medical Translation Research

2.1. Features of Medical Translation

2.1.1. More professional vocabulary. Medical English usually includes many scientific and technical meanings, all of which must have certain fixed conceptual meanings and attributes. One of the major differences between medical English terminology and everyday English terminology is that more than 65% of medical English terminology is directly obtained from Greek and Latin. For example, indigestion (indigestion) may be from Greek, chest may also be from Latin. The two language forms, pharmacology and Latin, which it needs to contain, are combined with other single nouns through their derivation and synthesis, thereby forming a large number of words used in pharmacology and Latin English.

2.1.2. Abbreviations are widely used. A large number of abbreviations are commonly used, which is an important feature of medical English. In order to be able to express and convey information more clearly, concisely, medical English generally uses abbreviations instead of original words. This abbreviation is often published in Western drug treatment prescriptions, laboratory reports and medical-related literature. The names of academic teams, medical journals and other medical units are usually words or abbreviations.

2.1.3. Noun structure frequently occurs. Chinese medical literature generally uses the form of verbs to describe and express the occurrence and appearance of sports, while Chinese medical English generally uses the structure of nouns, and the most important thing is to use a large number of nouns evolved from verbs. This can effectively reduce the production of some sentences or terms, but at the same time increase a lot of information.

2.1.4. Passive voice is used more often. In many medical literature, people initially only pay attention to the victim of one behavior, that is, the patient being studied, but few people pay attention to the behavior itself, that is, the medical service personnel. Therefore, the passive voice method is often studied and used in medical and English literature. The use of passive voice can express an event more objectively, avoid individual subjective emotions, and can provide predictive information and clear meaning.

2.1.5. Big sentences are very common. Parallel sentences and compound sentences are usually widely used in modern medical English to express complex logical relationships. Longer sentences generally contain more clauses and have a certain side-by-side structure. The non-predicate structure is missing or the opposite.

2.2. Advantages of Corpus-Assisted Translation

Translators can only refer to documents and reference books and dictionaries over and over again. They revised it several times and then wrote it word by word in the draft. This is a very
time-consuming task with high repetitiveness and high intensity, which brings great resistance to translation work. Computer-assisted translation system can help us to carry out translation activities faster and more convenient, which shows a very obvious technical advantage. The main advantages of computer-aided translation are as follows:

2.2.1. The ability to provide a large number of vocabulary translations. Since this translation system is actually a dictionary with a relatively small capacity, we can try to save the translation staff's time to request the text on the website. Computer-aided translation technology can apply professional knowledge to the grammar framework and vocabulary objects.

2.2.2. Provide full text translation. Using advanced translation and memory warehouse technology, the original sentences are automatically stored in a database. When used again, the search engine will be able to quickly analyze a large amount of data, quickly search and find out the materials closely related to the translated text in the shortest time.

2.2.3. Can provide as much information as possible about the translated content. A large amount of data can not only effectively improve the efficiency of language translation, but also effectively enhance the accuracy and consistency of language translation.

2.3. Auxiliary Medical Translation Improvement Strategy. Due to the complexity of multiple texts and rigorous grammar in modern English, it is difficult for robots to translate complex sentences correctly and only translate simple sentences. Secondly, because most English topics are sometimes ambiguous, our machines are challenging to explain automatically and can organize the correct language word order. For some high-quality translation work, translators can first use a computer to assist in translation, reduce the amount of work required before translation, and then use an editor to improve translation quality significantly. Before applying machine translation, you can make adjustments to the materials that need to be translated to adapt to the differences in Chinese and English expressions.

1) According to the logic of Chinese, break all big sentences, or subdivide them into simple sentences or short phrases that are easier to understand.

2) Make some adjustments to the sentence structure of English materials to adapt them to Chinese customs.

3) When translating, delete the untranslated links and general terms of the table.

4) By expounding in detail the passive medical language, basic structure and common grammatical features in the third-party English medical literature, input active speech to convert into noun-object structure. Finally, collect all the content, so even if the computer has translated a wrong language and sentence, we only need to change the language, and there is no need to find the exact language and errors.

2.4. Translation Search Algorithm

2.4.1. Machine translation method based on statistics. The model of machine translation can be simply understood as: For example, suppose a paragraph t is translated, but what is translated is a sentence s in the language, which is a process of being encoded; and for statistical machine decoding, it is Need to infer t from s to s, which is also a decoding process. Theoretically supporting formula is the Bayes formula

$$P(T|S) = \frac{P(S|T) * P(T)}{P(S)} \quad (1)$$

The translation process can be expressed as

$$T_{out} = arg \max \{ P(T|S) \} = \arg \max \{ P(T) * P(S|T) \} \quad (2)$$
2.4.2. Simhash algorithm. The specific implementation of the simhash algorithm can be divided into: first process the keywords in the original text and extract the keywords; then use a traditional hash function to calculate the fixed number of digits for each keyword (e.g. 32 bits); then for each hash value, if a certain bit is 1, then the fingerprint of the final text must be added to the value of the corresponding bit, otherwise it must be subtracted by 1, and finally the text fingerprint image obtained is carried out compression, the bits greater than 0 are compressed to 1, and the bits less than or equal to 0 are compressed to 0. Use sentences as text.

The algorithm estimates the cardinality based on the statistical information of the hash result. As the name implies, it is based on the minimum value of the sequence to estimate the cardinality. Assuming that the minimum value of the hash result of all elements in the multiset is X, the algorithm approximately considers the estimate of the set cardinality n to be one. The mathematical theory of the algorithm is: \([0,1]\) the minimum probability density \(f(x)\) of \(n\) random \(X\)-machine uniform variables in the interval is \(n(1-x)^{n-1}\), so the mathematical expectation of the minimum value satisfies the following equation:

\[
E(M) = \int_0^1 x \cdot n(n-1)^{n-1} dx = \frac{1}{n+1}. \tag{3}
\]

According to the calculation result of the minimum mathematical expectation of the interval \([0,1]\), it is easy to think: when the expectation of one is obtained, its value is approximately equal to \(M\). When the Simhash algorithm is analyzed and analyzed mathematically, it is found that \(x=0\) is mathematical Expect a divergence point of the integral equation, the following formula:

\[
E\left(\frac{1}{M}\right) = \int_0^1 \frac{1}{x} \cdot n(1-x)^{n-1} dx = +\infty \tag{4}
\]

3. Corpus-Assisted Medical Translation Experiment Based on Computer Technology

3.1. Experimental Design

Translation experiments were carried out under two different conditions using conventional reference resources (such as parallels and dictionaries). To facilitate the comparison between the experimental and control groups, please test as many variables as possible. The experiment takes 60 minutes. At the same time, in addition to the use of different reporting tools, the conditions in other aspects of the two groups are also exactly the same. Since the experiment is limited to the existing parallel resources (the main body's conditions will be explained in detail below), only the Chinese and English versions are selected for this experiment.

3.2. Subjects

The experiment subjects are English majors in the second semester of the third year of undergraduate colleges, 46 people. The subjects were randomly divided into two groups: experiment group (using corpus resources to assist translation) and the control group (using dictionaries and other conventional reference materials for translation). The experiment is scheduled towards the end of the term. At this time, the test students have carried out a certain amount of English-Chinese translation and Chinese-English translation practice, and have certain experience. Before the experiment, although the students learned some corpus concepts sporadically, and learned some basic monolingual corpus retrieval skills, they have never been exposed to the operation of parallel corpus, but they are more enthusiastic in using them and are willing to cooperate with parallel corpus to assist translation practice experiments. And received training on the operation and use of parallel corpus indexing software to ensure more proficient use.

3.3. Translation of the Original Source

This experiment uses the Lancet Chinese-English parallel corpus, collecting the original and English translations of the Lancet. The Chinese and English texts undergo sentence-level alignment processing to establish a connection between the original text and the target text. Chinese text is also processed by
word segmentation. Judging from the subject content involved in the text, the Chinese-English Lancet corpus used in this experiment is dedicated.

3.4. Experimental Environment
The hardware environment of the translation experiment is a networked classroom equipped with multimedia computers and Internet access functions. Students must download the translated original text, parallel corpus and retrieval tools from the online teaching platform as needed. After the experiment is completed, the translation will be submitted to the e-learning platform in the form of electronic text. In translation practice, all subjects obtain the original text through the Internet and submit the translation through the online teaching platform. The experiment did not change the students' translation environment for the experimental group, but added a new corpus tool.

4. Experimental Data Analysis

4.1. Amount of Translations Completed
A preliminary analysis of the translation was carried out based on the experimental data, and the data was sorted. The translation efficiency of the two groups was investigated from the amount of translation completed. The experimental results are shown in Table 1:

|                         | Test group | Control group |
|-------------------------|------------|---------------|
| Standard translation volume | 7231       | 7231          |
| Total translation completed | 4867       | 5981          |
| Percentage of completed translations | 67%         | 82%           |

It can be seen from Figure 1 that there is a significant difference in the total number of completed translations (total number of words) between the experimental group and the control group. The percentage of completed translation in the experimental group is 67%, while the percentage of completed translation in the experimental group is 82%. Thus, the efficiency is relatively low. This is
because the translation examples of the corpus query are vibrant, which increases the burden on the retrieval.

4.2. Translation Quality

According to the experimental data, the accuracy of term translation can be used to evaluate the translation quality of the translation. The experimental data is shown in Table 2:

|                         | Test group | Control group |
|-------------------------|------------|---------------|
| Number of terms to try to translate | 1390       | 1498          |
| Number of correct translation terms* | 1210       | 982           |
| Term translation accuracy rate      | 87%        | 65%           |

Table 2. Comparison of term translation accuracy

![Figure 2. Comparison of term translation accuracy](image)

It can be seen from Figure 2 that the accuracy rate of translation terms in the experimental group is 87% higher than that in the control group by 65%. Chi-square test further confirms that the experimental and translation terms have a clear advantage over the control group in terms of accuracy.

5. Conclusions

The medical parallel corpus helps translators to conduct in-depth research and thinking about the problems encountered in translation, which may improve the translation quality. Still, under current technical conditions, it doesn't seem easy to improve the efficiency of translation work by querying the Ping corpus. This is especially true for those learners who rely too much on corpus queries and frequently use corpus queries in the translation process. There are discrepancies between this conclusion and the conclusions in the relevant literature. However, the experiment in this paper verifies that the parallel corpus has the effect of optimizing translation.

Acknowledgment

Research Program on Humanities and Social Science of Education Department of Shaanxi Provincial Government (21JK0343);
“English Language and Literature”, the 4th Key Construction Program of Xi’An Medical University ([2019]96); Counterpart Supporting Program of Xi’An Medical University (2018PT51, 2018PT52); Education and Teaching Reform Program of Xi’An Medical University (2018JG-48, 2018JG-51); Innovation and Entrepreneurship Program of Education Department of Shaanxi Province (S202011840043, S202011840029); Enterprise-University Cooperative Education Program of Ministry of Education (202002268025).

References
[1] Coley C W, Rogers L, Green W H, et al.(2017) Computer-Assisted Retrosynthesis Based on Molecular Similarity[J]. ACS Central Science, 3(12):1237-1245.
[2] Zhang W.(2017) Research on the innovation of English vocabulary teaching based on computer-assisted corpus[J]. Revista de la Facultad de Ingenieria, 32(2):550-556.
[3] Tholpadi G, Bhattacharyya C, Shevade S.(2017) Corpus-Based Translation Induction in Indian Languages Using Auxiliary Language Corpora from Wikipedia[J]. ACM transactions on Asian language information processing, 16(3):20.1-20.25.
[4] Scandura, Gabriela.(2016) Corpus-based Translation and Interpreting Studies: From description to application/Estudios traductológicos basados en corpus: de la descripción a la aplicación[J]. 62(4):688-692.
[5] Wenchao, Defeng.(2016) Corpus-Based Studies of Translational Chinese in English-Chinese Translation[J]. Literary & linguistic computing: Journal of the Alliance of Digital Humanities Organizations, 31(3):516-519.
[6] Malmkjir K.(2018) Children’s Literature in/and Translation: The Oeuvre as Corpus[J]. Ilha do Desterro A Journal of English Language Literatures in English and Cultural Studies, 71(1):135-150.
[7] Caimotto M C, Gaspari F.(2018) Corpus-based study of news translation: Challenges and possibilities[J]. Across Languages and Cultures, 19(2):205-220.
[8] Grabowski L.(2020) Phrase Frames as an Exploratory Tool for Studying English-to-Polish Translation Patterns: A Descriptive Corpus-Based Study[J]. Across Languages and Cultures, 21(2):217-240.
[9] Su W, Li D.(2016) Corpus-Based Studies of Translational Chinese in English-Chinese Translation[J]. Literary & linguistic computing, 31(3):516-519.
[10] Matvieieva S.(2020) Terminological Synonymy: Corpus-Driven Translation[J]. International Journal of Science and Research (IJSR), 9(3):971-973.
[11] Li Y, Ren Y.(2016) Database Design on Corpus System for Chinese-English Translation of Scientific Papers[J]. International Journal of Simulation: Systems, 17(13):15.1-15.7.
[12] Aliye, Abdughini, Aliya, et al.(2018) Corpus-based approaches to translation and interpreting: from theory to applications[J]. Perspectives: studies in translateology, 26(1-2):153-156.