Hindawi
Scientific Programming
Volume 2022, Article ID 2264235, 10 pages
https://doi.org/10.1155/2022/2264235

Research Article

Construction of Medical Academic English Translation Model Driven by Bilingual Corpus-Based Data

Jinping Liu [1] **and Hong Liu** [2]

[1]*Department of Foreign Languages, Xinyang College, Xinyang, Henan 464000, China*
[2]*Department of Foreign Languages, Henan University of Science and Technology, Luoyang, Henan 471000, China*

Correspondence should be addressed to Hong Liu; 9902652@haust.edu.cn

Received 16 March 2022; Revised 11 April 2022; Accepted 15 April 2022; Published 9 May 2022

Academic Editor: Sheng Bin

With the rapid development of information collection technology and natural language processing technology, the construction of English-Chinese bilingual parallel corpus has developed rapidly. The scale of corpus and related technology have a large research space; and how to obtain effective data and knowledge from massive resources, in order to better serve the basic and applied research, is becoming a trend. Based on the research of parallel bilingual corpora at home and abroad, this article extracts the text features of medical English and constructs English-Chinese bilingual corpora with different levels aligned. Based on the statistical analysis of the distribution characteristics of phrase structures and the acquisition of characteristic knowledge, an English-Chinese bilingual translation model is constructed based on the bilingual corpus, and then the phrase structure knowledge in English and Chinese sentences is extracted from the data-driven English-Chinese bilingual corpus through the model. The results show that the accuracy of the test set is 91.63% and the F value is 90.05% under the condition of keeping the recall rate stable. The accuracy of the text structure features has been significantly improved, and the expected effect has been achieved through the test.

1. Introduction

With the rapid development of science and technology, bilingual corpus plays an increasingly important role as a basic resource, and the construction of bilingual parallel corpus also gradually increases [1]. Bilingual corpus has gradually become the main content of language research, and bilingual corpus combined with relevant software has achieved fruitful results in bilingual translation research and bilingual lexicography [2]. The construction of bilingual corpus contains corresponding relations at different levels of text, paragraph, and sentence of two languages, which provides the research foundation for many fields of natural language processing [3]. In the era of big data, information is characterized by diversity and real-time performance. With the development of technology, a variety of corpora have emerged. Today's information processing systems are inseparable from the support of data and knowledge base. As the basic language database and knowledge base, corpus has formed the basis for the realization of natural language processing methods at different levels. Bilingual corpus is a kind of corpus that contains the information of translation between two languages. It can provide rich matching information between two languages and has important application value in instance-based and statistics-based machine translation and word sense disambiguation. The construction of bilingual corpus, the accuracy of sentence alignment and block alignment are of great importance to the quality of machine translation, and corpus is an indispensable resource in translation system. A translation system relies on available training data; and the more data are available, the better the parameters of the translation model can be estimated. The closer the model is to the possibility of real translation, the higher the translation performance is. [4]. Therefore, relying only on artificial way to carry out information processing will not be able to meet the current

practical needs. The construction of English-Chinese bilingual corpus plays an important role in promoting medical science and technology. The research of bilingual corpus construction is becoming more and more urgent and has gradually developed into a research hotspot in the field of natural language processing.

Medical English is a branch of English for Science and technology. The characteristics of English for Science and Technology are focusing on narration, logical coherence, clarity and fluency in expression; avoiding ambiguity in writing and avoiding strong personal feelings and subjective arbitrariness in argumentation; using less descriptive adjectives, lyrical adverbs, and interjections; paying attention to simplicity and precision; trying to avoid using all kinds of rhetoric aimed at strengthening the language appeal and propaganda effect; and avoiding exaggeration and sarcasm. Irony and other rhetorical means to make readers feel pompous and distorted [5]. Medical English not only conforms to the above characteristics, but also has a high degree of concept, abstractness, objectivity, and reasoning. The semantic expression should be accurate, the structure should be rigorous, the level should be clear and the logic should be rigorous, and the terms should be professional, written and international universal. Therefore, to understand and master the stylistic characteristics of medical English from the perspective of scientific stylistic is of great guiding significance to the practice of medical English translation. The topics of medical English are rich and varied, including drug instructions, medical records, medical papers, medical reviews, medical record reports, clinical trial plans, and popular science articles. Articles with different themes have different writing purposes and reader groups and have their own characteristics in vocabulary, grammar, and style, presenting a high degree of "textual theme specialization" [6]. That is, similar themes have roughly the same format requirements and expression methods. But generally speaking, the main characteristics of medical English are familiar, mainly reflected in vocabulary and sentence. Medicine is a highly specialized and scientific discipline, so the biggest characteristic of medical English is a large number of medical professional terms, which are almost never seen in articles in other fields. Moreover, vocabulary is the basis of language, so to understand medical English, we must first understand medical terms. Understanding the composition and characteristics of medical vocabulary plays an important role in correctly understanding and translating medical English literature.

2. Related Work

The construction of bilingual corpus is a new trend in the horizontal development of linguistics. Language researchers have clearly recognized the great role of high-quality and large-scale bilingual corpus in language comparative research and language teaching. It has been more than 20 years since the construction of bilingual corpus in foreign countries, and a lot of research has been carried out based on bilingual parallel corpus, but mainly on the translation research and language comparison between Spanish. In China, many scholars in linguistics and translation have also built large and small bilingual parallel corpora since the late 1990s. Chen et al. constructed four parts of the corpus, including encyclopedia corpus, translated text database, translation statement database, and specialized corpus and realized the sentence-level alignment of Chinese-English texts, which can be used for basic grammar annotation and automatic link retrieval of word frequency, phrase, sentence pattern, and collocation. The same interface realizes the bidirectional alignment of English and Chinese sentences. Empirical research on the basis of corpus mainly includes translation language features or translation commonalities, translation units, translation texts, and translation teaching. At present, a Chinese-English bilingual online retrieval platform of about 10 million words has been built, providing references for language and translation researchers [7]. The English-Chinese parallel corpus developed by Chen et al. adopts the principle of sentence alignment as the primary and paragraph alignment as the secondary and is used in Chinese-English comparative studies and bilingual lexicography [8]. The constructed parallel corpus has two characteristics: on the one hand, sentence alignment is realized, which is convenient for bilingual comparison and bilingual translation research; and on the other hand, the combination of traditional method annotation and automatic annotation makes relevant studies further [9]. In order to improve the bilingual alignment effect, Gu divided the text-aligned files into paragraph-aligned files according to the characteristics of the data [10]. The block-based translation method is conducive to solving ambiguity in machine translation. However, all sentences are analyzed into different types of blocks. The current technology is not able to control countless blocks, and it is even difficult to determine the boundary of blocks. Candel Mora proposed a large-scale Chinese-English bilingual parallel corpus construction system and realized an automatic acquisition system for bilingual corpus mining through content analysis and link analysis of web pages by utilizing massive multilanguage text resources on the Internet [11]. Xu proposed the construction of a large-scale Chinese-English bilingual parallel corpus. The parallel corpus of English sentence alignment is mined from Wikipedia, which is a method to extract sentence-level alignment by using the bilingual dictionary method based on extended link [12]. Curry and Chambers put forward a kind of using multiple features through the k-neighbour classifier to identify the parallel texts and automatic data collection from the Internet high-quality bilingual parallel corpus, the method of high translation accuracy in both English and Chinese parallel corpus acquisition methods, including from online news, online dictionary, and translation site to get the data and the parallel aligned text in a sentence or a document level [13]. Li and Pan et al. proposed an automatic Chinese and English parallel corpus construction system, which includes word-level alignment and character-level alignment and uses the longest common subsequence to find the most reliable Chinese translation of English words to build a bilingual corpus [14]. English model can help scholars translate papers quickly and greatly reduce the workload. Medical academic translation model is of great significance.

This article draws lessons from the bilingual parallel corpus method of researchers at home and abroad, extracts text features of professional words, uses abbreviations to make reading easier, and constructs English-Chinese bilingual corpus with different levels of alignment to help scholars quickly translate academic papers.

3. A Bilingual Corpus Data-Driven Model for Medical Academic English Translation

3.1. Extract Text Features of Medical English. In this article, there are about 50000 basic vocabularies of medical profession. The derivations are composed of a ready-made word or root plus some morphemes that do not exist alone but have fixed meanings. It has the characteristics of simplicity and accuracy of meaning, word formation shows the meaning of words, segmented formation is easy to remember, and root and affixes have great potential [15]. It is through this derivation that medical English has a large number of words that can express the rich, subtle, and complex medical science and can constantly produce new words to meet current and future needs with the development of medical science and technology [16]. Since derivations are morphemes formed according to certain rules, and they are polysyllabic and polymorphic, it is usually possible to infer the meaning of a word once the individual components are known. An abbreviation is a brief form of a word or phrase. It has the characteristics of being professional, informative, concise, and convenient. There are many acronyms in medical English. Because medical terminology is long and complex, an acronym is often created to condense common information in order to facilitate the exchange of medical information. There are a large number of long professional words in medical English literature. Using abbreviations can make reading easier to understand. People in the industry often use such abbreviations in oral communication and literature. The formation of abbreviation is as follows: first, it is composed of the first letter of each content word. For example, most of the examples mentioned above are composed of the first letter of each content word. Second, extract the main letters in the original words and phrases. In medical English, some diseases, scientific research methods, anatomical structures, and operations are named after people or place names [17]. Generally speaking, most of the named terms are named after people, most of which are named after the last name of the discoverer or creator of new principles, new methods, anatomical structures, symptoms, and diseases.

3.2. Extract Syntactic Features of Medical English. Due to the complexity of the object described in medical English and its logical and rigorous expression of the structure, it is inevitable to use a large number of long and complex sentences to fully express and emphasize complex and accurate information. Medical statement accuracy and rigor are closely related to disease and its corresponding treatment, so the medical English sentences must pass all the information on a particular point, including the level of the patients' health status, disease and virus, treatment conditions and requirements, etc., as any omission may cause the reader to misconception and misunderstanding of space also it will lead to treatment errors and even endanger patients' lives [18]. Passive voice is widely used in scientific English, especially medical English. The passive voice focuses the reader's attention on the thing, phenomenon, or process being described because it better highlights the object to be explained and argued about, namely the disease or patient. The passive voice conforms to the right-branch syntactic structure of English. The passive voice has more room for structure and avoids mentioning medical personnel. It is conducive to highlighting important concepts, problems, facts, conclusions, and other contents and is more conducive to expanding noun phrases and information content, which plays a major role in controlling information flow [19]. The passive voice avoids the personal pronoun and reduces the subjective colour, thus achieving the goal of paying attention to objective facts in medical literature. The passive voice makes the sentence structure more concise and compact because of its structural characteristics and also makes the sentence varied and not rigid.

3.3. Construction of English-Chinese Bilingual Corpus. When making an Excel database in Both Chinese and English, 50000 basic words of medical profession were collected and inputted. According to different research purposes and uses of corpus, that is, different principles and methods of corpus collection, corpus can be divided into four types: heterogeneous, homogeneous, systematic, and special. According to the language of the corpus, the corpus can also be divided into monolingual, bilingual, and multilingual types. This article focuses on the data-driven construction of English-Chinese bilingual corpus, including corpus design, corpus collection, and corpus processing. Bilingual corpus is a special corpus containing the information of translation between two languages, which can provide rich matching information between two languages. The specific process of English-Chinese bilingual corpus construction is shown in Figure 1.

Alignment technology is the core of bilingual text processing. The so-called alignment is the process of finding translation fragments from the translated texts of different languages. The alignment of English-Chinese bilingual corpus can be divided into different levels of processing depth, such as paragraphs, sentences, and chunks. The processing depth of English-Chinese corpus determines the granularity of knowledge provided by the corpus. Bilingual alignment is to establish the corresponding relationship between the same language units of the source language and the target language in the bilingual corpus, that is, to determine which language units in the source language text and which language units in the target language text are mutually translated. Bilingual text has multilevel and multigranularity correspondence, including the alignment between paragraphs, sentences, and blocks. As corpus contains a large amount of corpus, it is undoubtedly inefficient to rely

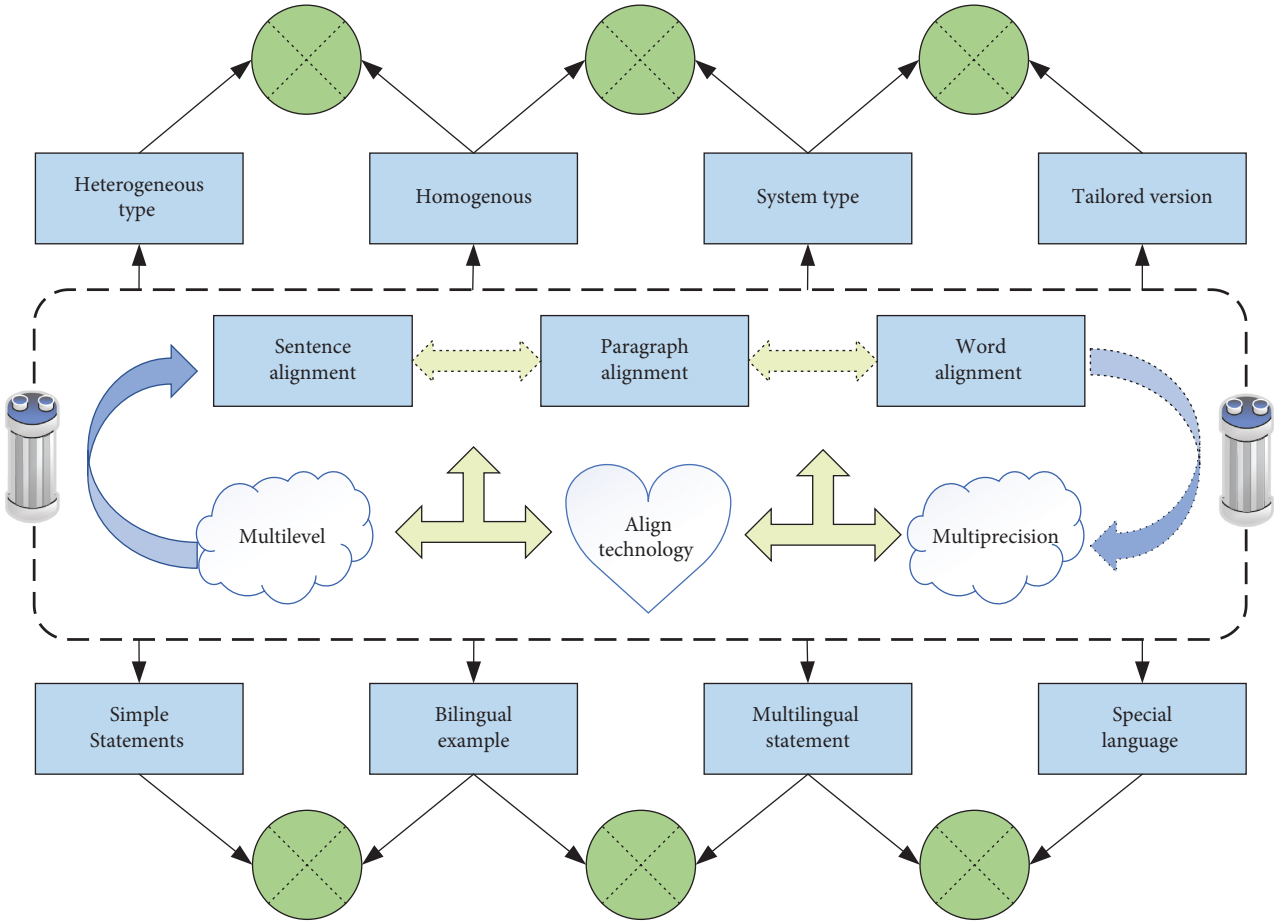

Figure 1: The flow chart of English-Chinese bilingual corpus construction.

only on manual methods for such alignment, so automatic alignment technology emerges. In bilingual corpus, sentence alignment is the key, which plays a connecting role between paragraphs, sentences, and chunks. In order to construct a certain scale bilingual corpus, the bilingual data are processed into English-Chinese sentence alignment parallel corpus through the bilingual sentence alignment platform.

3.4. The Overall Structure of Medical Academic Translation Model. The medical academic translation system as a whole is mainly divided into two modules: data support and translation system. Data supports mainly includes data mining and data cleaning. In the early stage, a large amount of bilingual corpus needed for medical academic translation should be mined, and the scale of data has a great influence on the learning performance of the translation model. Not all raw data are purely directly available, so we need further data processing operations. In order to eliminate noise pollution in large-scale data, it needs to be processed according to some cleaning requirements of the specific language feature corpus. The translation system mainly involves data preprocessing, vocabulary pretraining, and translation model training and testing. The technical structure involved in translation model training is very complex [20]. To improve the overall performance of translation model, a thorough understanding of the internal structure and experimental verification are required. In this article, the overall process of the model is first introduced. The input text to be translated is preprocessed, including word segmentation, clause, segmentation, part-of-speech tagging, information extraction, etc. The constructed hierarchy is used to carry out corresponding text representation and topic clustering for extracted text features to be translated. The medical translation model is used to translate the text, and the translation quality analysis module is used to analyze the quality of the translated text. The overall architecture of the medical academic translation system is shown in Figure 2.

In the field of natural language, preprocessing is a basic step in various text analysis tasks and plays an important role in subsequent semantic analysis. Using natural language processing tools, the preprocessing of English composition mainly includes special character filtering, segmentation, clause, word segmentation, word suspension and word stem processing, part-of-speech tagging, and information extraction. Part-of-speech tagging is the process of tagging words in English compositions according to their context and definition. In this article, a part-of-speech tagger will be used to assign a part-of-speech tag to each word of the segmented English text. Part-of-speech tagging is very necessary for the construction of dependency syntax tree.

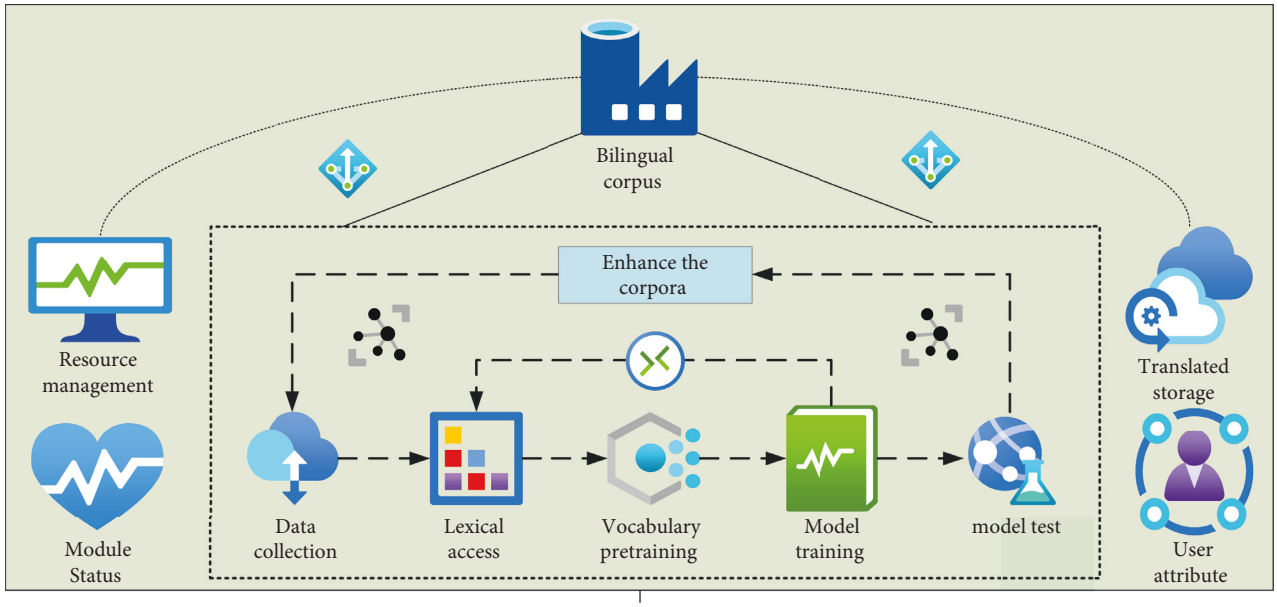

Figure 2: The overall architecture diagram of medical academic translation system.

Rule-based part-of-speech tagging combined with lexica-based part-of-speech tagging can tag the words that do not exist in the training corpus but exist in the test set. Based on probability, part-of-speech tagging is carried out for a specific tag sequence depending on the probability of occurrence. It can mark the parts of speech of words by analyzing global and local semantic information, and the accuracy of part-of-speech tagging is relatively ideal.

3.5. Structural Design of Translation Model for Bilingual Corpus. An integrated system is designed based on architecture model analysis, which can complete translation work including data preprocessing, word vector training, model training, and model testing. Training can save the optimal performance of n models. Finally, these models are fused and evaluated. Such a system structure will reduce the labour cost. The whole training test result is very simple. We only need to load the double corpus catalogue we have sorted out and run the script, so we no longer need too much labour cost to test the value of each model in real time and finally get the result. The architecture of bilingual corpus translation system is shown in Figure 3.

Before the training, the parallel training corpus is loaded first, and the model pretrains the word vector of the parallel corpus to obtain the word vector of source language and target language, respectively. Then all the corpus pairs in the form of these word vector are inputted into the model for training. Before the training, a model saving frequency will be set first. During the training, the iteration of saving frequency has not gone through, and the error value needs to be checked to determine whether it is within the set range. If so, the model will be saved and then tested. Training cannot go on all the time. When our training iterations or training times exceed the set threshold value or the accumulated error value reaches the present threshold value, then our training gets ended.

4. Experiment and Analysis of the Model

This paper uses a high operation and maintenance performance platform for model construction and training. In the training phase of the model, the word embedding model needs to use a large number of training corpora to distribute word vectors. Since the semantic information of the corpus is time consuming, in order to effectively shorten the training time of the model, this paper uses a laboratory workstation for training in the process of training the model.

4.1. Translate Model Experimental Data Sets and Evaluation Criteria. The experiment trained the relevant training set needed to construct bilingual corpus. The training sets used in this article include Chinese Learners corpus, Asian English Learners International Corpus, and Wikipedia corpus. The medical academic English translation model constructed in this article is trained and tuned, and the optimal parameters set in this article are adjusted through the training sets. In order to verify the applicability of this model to students in other English-speaking countries, part of the data set of foreign students is selected as a test set. Based on the selection principle of the test set above, this article selects Chinese students' medical academic papers on 5 topics from the Chinese Learners corpus. In this article, based on the principle of topic differentiation, a large number of composition samples of subject categories are selected from the test set. The above test sets will be used to comprehensively verify the validity and accuracy of the analysis method in this model.

The evaluation criteria of the model experimental results in this article are accuracy, recall rate, and F value, which are widely used in NLP analysis and machine learning.

The accuracy of the model refers to the ratio between the number of samples with positive predicted samples and the number of all positive evaluated samples in the text that the

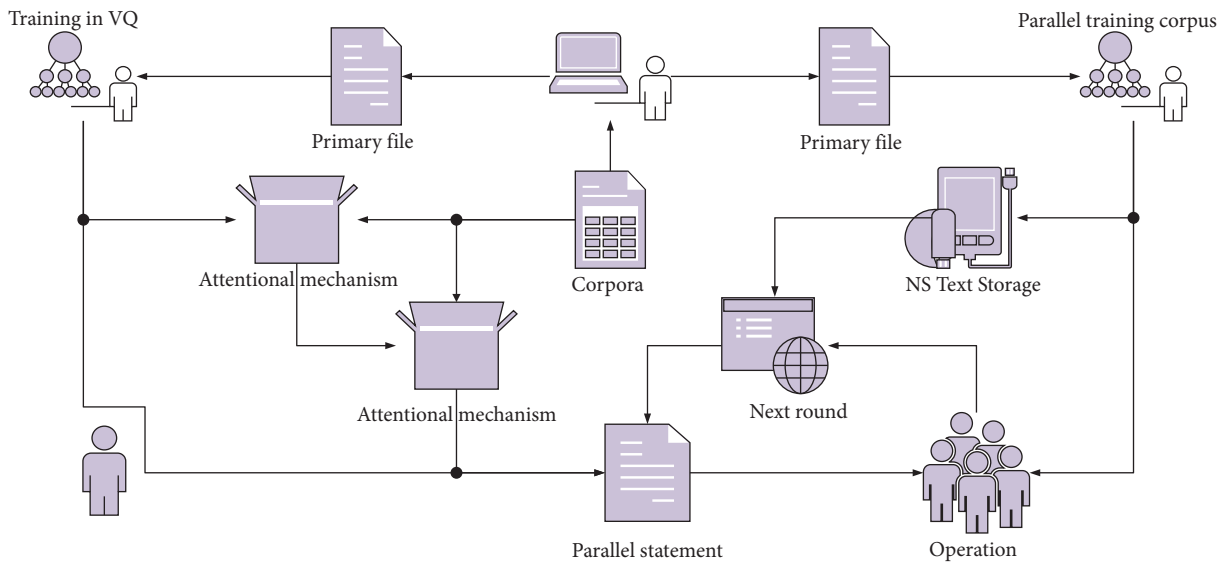

Figure 3: Architecture diagram of bilingual corpus translation system.

machine evaluation results are consistent with the manual evaluation results [21]. The higher the ratio is, the higher the accuracy is, reflecting the more ideal the effect of machine evaluation.

$$P = \frac{A}{A + B}. \tag{1}$$

The recall rate of the model represents the ratio of the number of positive samples that are consistent with the results of machine evaluation and manual evaluation to the number of all samples that are positive in manual evaluation, which is a measure reflecting the coverage of machine evaluation. The higher the ratio is, the higher the recall rate of the model is.

$$R = \frac{A}{A + C}. \tag{2}$$

The F value of the model is the weighted harmonic average of accuracy and recall rate. The larger the F value is, the more ideal the effect of text classifier is.

$$F = \frac{2 \times P \times R}{P + R}. \tag{3}$$

4.2. Evaluation of Medical Academic English Model. This article is based on a bilingual corpus-driven model test of medical academic English translation. The translation model has text in both languages, and the text in both languages describes the same information. Therefore, it is different from using translation model to construct corpus of resources. The model with bilingual corpus is a hybrid model. The method is usually divided into two steps: the acquisition of bilingual mixed corpus and the extraction of translated texts. English parallel sentence pairs can be obtained from bilingual mixed corpus. Then, the model is used to point out the characteristics of other bilingual models. By searching for the text pairs with translation words in the corpus, the existing form of bilingual text is abstracted by using the text pairs. These templates are then applied to the entire website model to extract text with the template. The test set of the experiment consisted of 14,000 academic papers selected from 10 topics in different corpora. In order to verify the performance of the thesis test sets based on bilingual corpus in different corpora, experiments were carried out in the thesis test sets of different topics in different corpora. The experimental results are shown in Figure 4, which reflects the accuracy, recall rate, and F value of the model in different composition test sets.

As can be seen from the figure, the translation model has an ideal effect under the test set of medical academic papers with different topics and different corpora, and the average F value of 10 medical academic papers is 92.04%. From the use of medical academic papers of varying lengths and test set of test results, medical academic papers' corpus test set accuracy is higher than the accuracy of the data set foreign English composition. One of the reasons is that we use Chinese students' medical academic papers as the training set for training. The parameter setting is more suitable for the translation of medical academic papers. Another part may be that the topic of foreign medical academic papers is relatively open, and the semantic information of the subject is relatively scattered. Therefore, there is a small difference in the experimental results of the test set. As can be seen from the broken line chart of the experimental results, under the condition of keeping the recall rate stable, the accuracy of tangential analysis of English compositions in the test set is 91.63% and the F value is 90.05%, showing good experimental results.

4.3. Model Loss Function. In order to test the classification effect of the classifier, 2000 pairs of web pages were randomly selected and manually marked whether the web pages were translated or not. The statistical analysis method of cross-validation is used to test the performance of the classifier. The specific methods are as follows: The 2000 manually

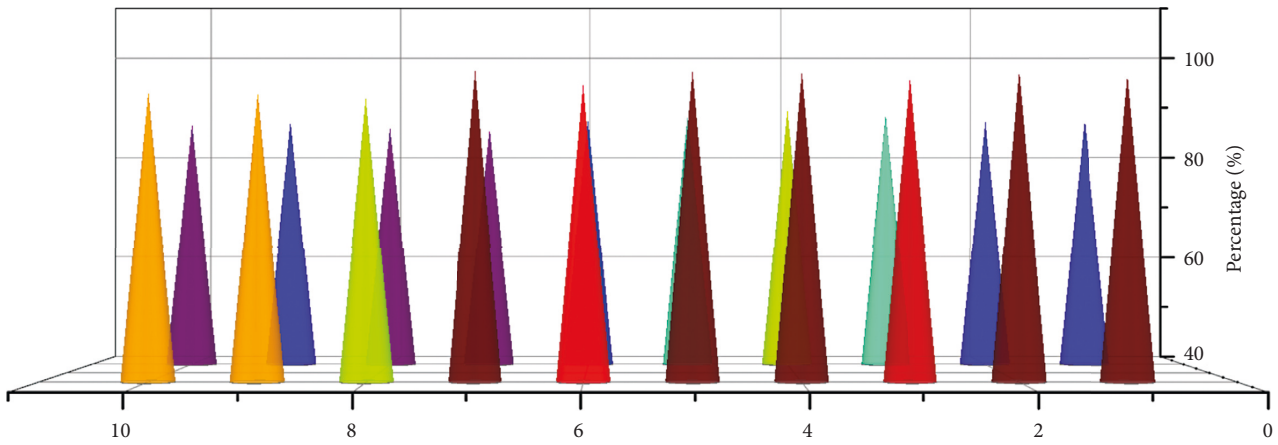

Figure 4: Experimental results of medical academic English model evaluation.

labelled samples were divided, 80% of which were used as the classifier training set and 20% as the test set. After a long time of training, the loss function of the model decreased with the increase of iterations in the training process. In the process of model training, the learning rate adjustment strategy of learning rate rising first and then falling was adopted. As the model was trained, the schematic diagram of its learning rate was shown in Figure 5. The main idea of the evaluation index is that the closer the results of machine translation are to those of manual translation, the better the machine translation model will be.

The experimental results show that the performance of the model is better than that of the basic model, and the correlation between the source language and the target language is improved during model calculation, so as to improve the performance of the model. If the score of all categories except the correct category is added up and if the score of the correct category is 1 higher than the score of the wrong category, the loss function of this image is 0. If the wrong category score is higher than the correct category score, the loss function is the difference between the two, which is also called the hinge loss function. X-axis represents the score obtained from the classifier on the correct category label of the text in the training set, and Y-axis represents loss. As Y increases, the loss function gets smaller, and it goes to zero beyond a certain threshold. The training model uses a large amount of unlabelled data to learn language knowledge from large-scale data, which enables the training model to capture more general language laws and has strong coding and language interpretation ability. The model can learn more grammatical representations of the source language by feeding sentence vectors containing the parts of speech, syntax, grammar, and semantics of the source language into the model for training. The accuracy of text structure feature classifier is improved obviously, but the recall rate is decreased slightly. Using all features, the performance of the classifier is optimal.

4.4. Comparison between Model and Human Translation. Extracting knowledge from unstructured text, such as keywords, terms, and chunks, has always been one of the important research contents in natural language processing and text mining. This article attempts to extract English and Chinese phrasal knowledge from English and Chinese bilingual sentence pairs using machine learning based on bilingual parallel corpus. Due to the complex types of phrase structures, it is impossible to extract all the phrase structures one by one in the limited content of the article. In this article, the representative phrase structure is selected to extract the phrase structure, hoping to provide a reference method for the extraction of the whole phrase structure. The length and internal structure distribution of phrase structure have prominent distribution characteristics, which can make it have the characteristics of short and simple phrase structure and take into account the characteristics of long and complex phrase structure in the extraction process. On the basis of corpus selection, the syntactic function structure is used as 150 syntactic structures. According to the structure label in the corpus, the English phrase structure in its parent node is counted as a syntactic function. Specific English phrase structures appear in the top ten of the relatively high frequency distribution of syntactic structures, as shown in Figure 6.

As can be seen from the figure, the parts of speech of Chinese structured unary adjacent are mainly concentrated on verbs, startles, and adverbs, which cover 80.93% of the vocabulary. If the words covered by nouns and auxiliary words are also counted, the five parts of speech contain 89.03% of unary adjacent. The above data indicate that the part-of-speech feature knowledge, especially the five part-of-speech feature knowledge, should be utilized in the process of training structure recognition model through conditional random field. The parts of speech of English phrasal structure unary adjacent are mainly concentrated on nouns such as NN, NNP, NNS, and NNPS, occupying 81.71% of the

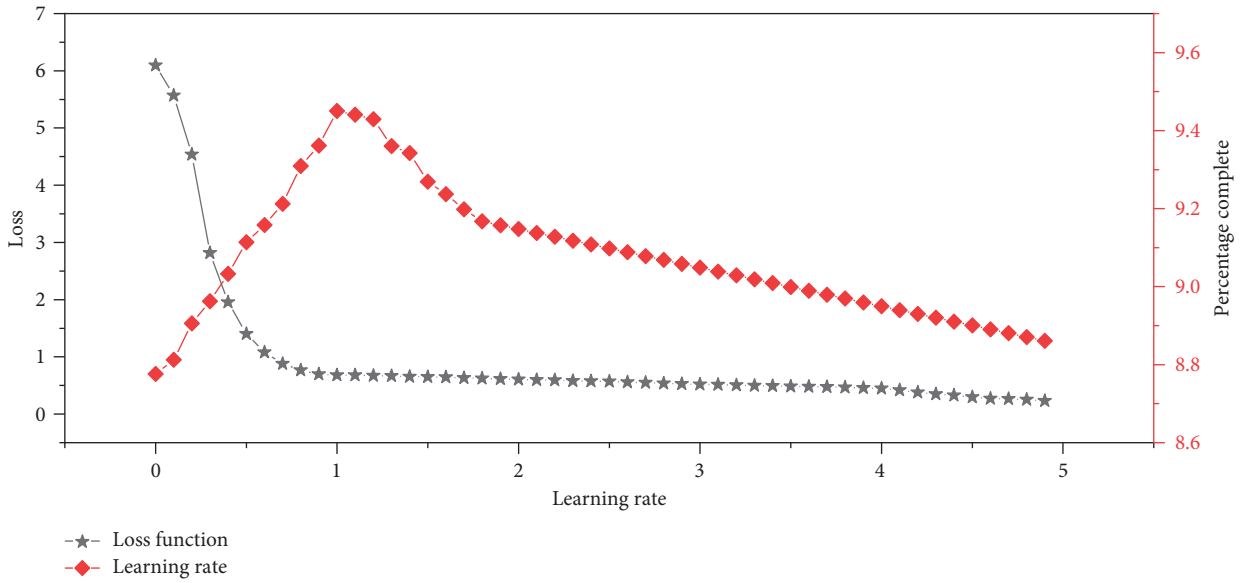

Figure 5: Schematic diagram of model learning rate.

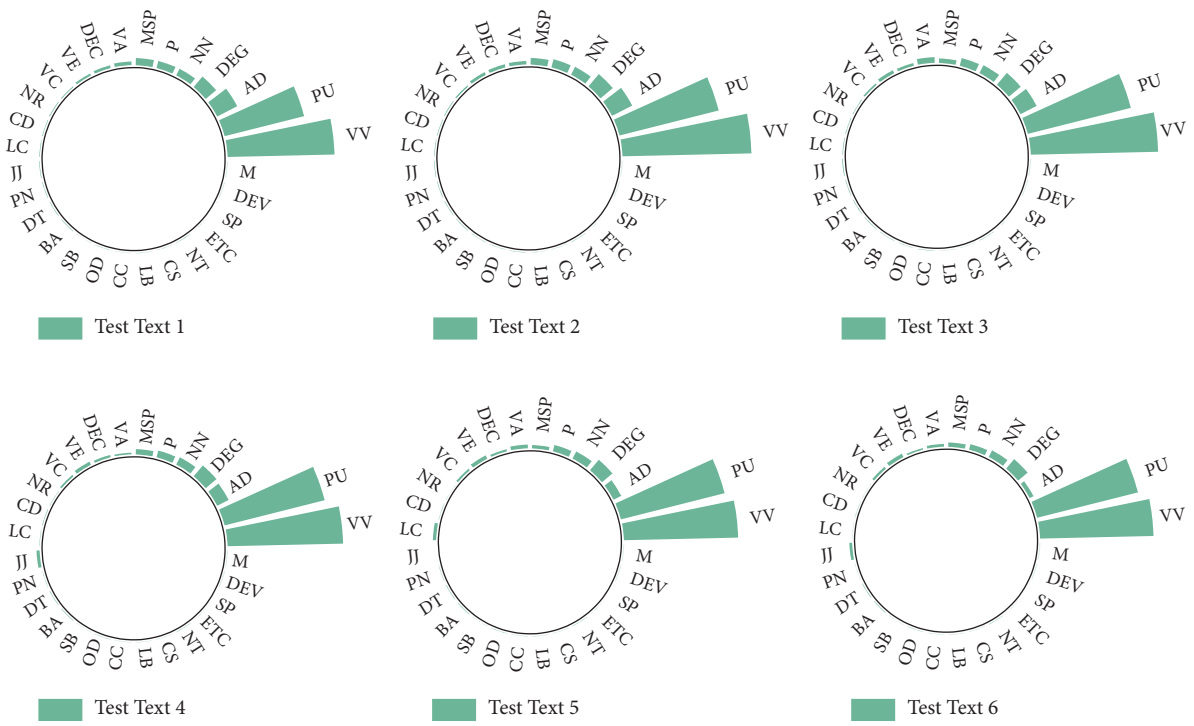

Figure 6: Quality score experimental results diagram.

whole parts of speech; compared with Chinese, the distribution of English phrases is relatively concentrated. In addition, it is numerals, accounting for 6.28% of the whole parts of speech. The total number of verbs is only 1.72%, which is very different from the 40.94% of verbs in the right boundary of Chinese phrase structure. 31.41% of the right boundary of Chinese phrase structure is punctuation, while only 1.75% of the right boundary of English phrase structure is punctuation. This shows that the object phrase structure of English is concentrated, while that of Chinese is discrete. This feature of English parts of speech indicates that in the construction of English phrase structure knowledge acquisition model, the corresponding knowledge of noun class should be fully explored and utilized to improve the performance of the whole model.

4.5. Comparison between Model and Human Translation. In order to verify the effectiveness of this model in English-Chinese translation, the bilingual corpus is used to translate

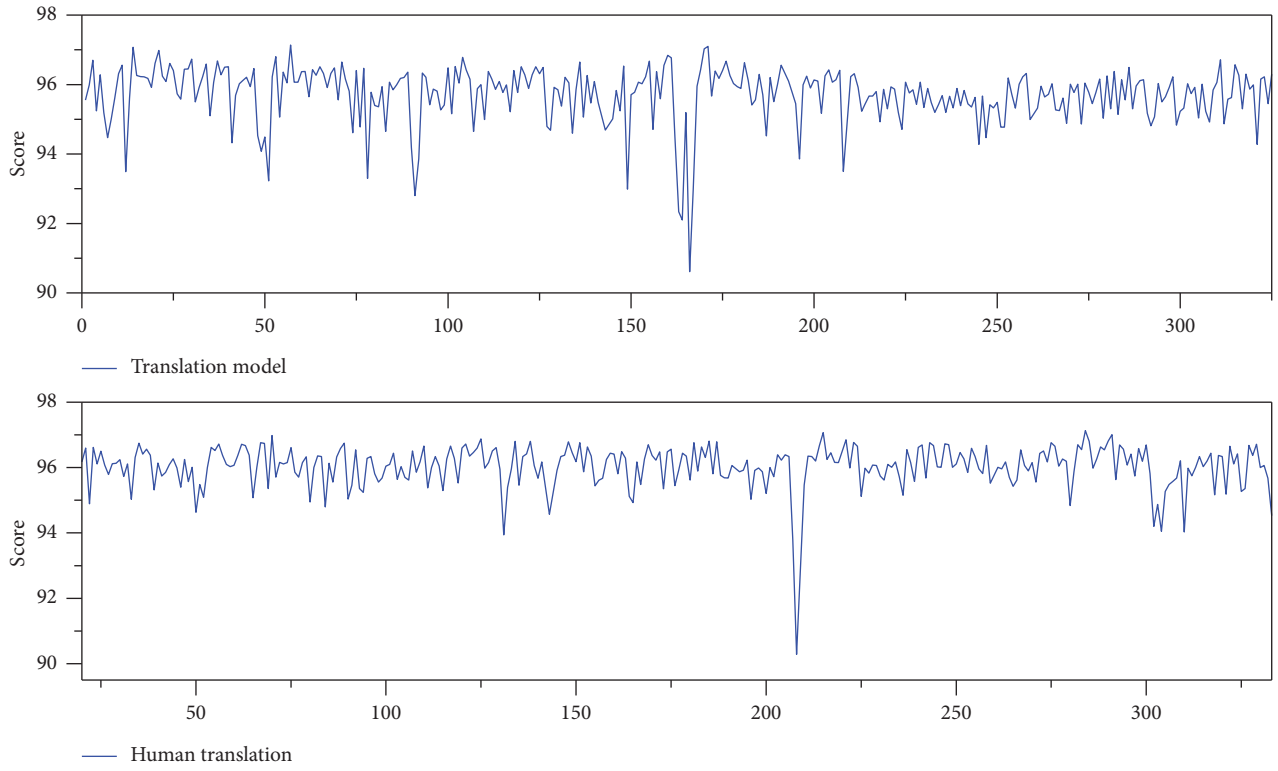

Figure 7: Quality score experimental results diagram.

medical academic papers according to abstractable steps, text pair acquisition, and parallel corpus extraction as it does not have obvious external features of web pages. In this step, the keyword set of the original web page is extracted, and then the keywords in the keyword set are translated into the translated keywords. Then the target documents containing these keywords can be retrieved by using these translation keywords. Then, these documents are sorted and filtered and the most matching translated web documents and original documents to form a group of text pairs are selected. Some mature language-specific sentence alignment tools are often used for sentence segmentation and sentence alignment between original and translated documents. We invited a number of professional English tutors in our research group to score the quality of 50 selected English composition test sets and compared them with the quality score of this model. The specific experimental results are shown in Figure 7.

From the analysis and comparison experiment of 50 medical academic translation papers, it can be seen that the score of the model in this article is close to that of teachers on the samples of medical academic translation papers in most test sets. The average score of teachers in the test set of medical academic translation papers is 81.73 points, and the average score of this model is 83.64 points. The difference between teachers' score and this model's average score is 1.91 points. By evaluating the comparative experiments in different test sets and analysing the comparative experiments in medical academic translation papers, the medical academic translation model has high accuracy and provides more detailed analysis of the content of medical academic translation papers by injecting professional words and modifying incoherent sentences, which has a relatively ideal effect.

5. Conclusion

Translation is a bridge for people who use different languages to communicate. The content expressed in medical English is professional, precise, and standardized. To do a good job in medical English translation, it is necessary to have a high English language application ability. This article studies the unique linguistic features of medical academic English, extracts the textual features of medical English, and constructs the English-Chinese bilingual corpus with different levels of alignment. In order to improve the performance of the translation model, the knowledge of phrase structure in English and Chinese sentences is extracted on the premise of statistical analysis of the distribution characteristics of phrase structure and acquisition of feature knowledge. A large-scale English-Chinese bilingual sentence alignment corpus is used to train the translation model, and sentence vectors containing information such as parts of speech, syntax, grammar, and semantics of the source language are fed into the model for training, so that the model can learn more grammatical representations of the source language during the training process. The accuracy of text structure feature classifier is improved obviously, but the recall rate is decreased slightly. The object phrase structure of

medical academic English is centralized. In the construction of English phrase structure knowledge acquisition model, the corresponding knowledge of noun class should be fully explored and utilized to improve the performance of the whole model. Due to the limitation of time and level, the quality of corpus in bilingual hybrid websites and cross-site bilingual pages is relatively better, but it is difficult to mine. In the follow-up work, we will continue to study how to divide the acquired corpus into domains or acquire the corpus by domain.

Data Availability

The 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 known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Acknowledgments

This study was supported by the Department of Foreign Languages, Xinyang College.

References

[1] J. Liu and L. Han, "A corpus-based environmental academic word list building and its validity test," *English for Specific Purposes*, vol. 39, pp. 1–11, 2015.

[2] M. V. Farahani and R. Kazemian, "Speaker-audience interaction in spoken political discourse: a contrastive parallel corpus-based study of English-Persian translation of metadiscourse features in ted talks," *Corpus Pragmatics*, vol. 5, no. 2, pp. 271–298, 2021.

[3] Ł. Grabowski, "Phrase frames as an exploratory tool for studying English-to-Polish translation patterns: a descriptive corpus-based study," *Across Languages and Cultures*, vol. 21, no. 2, pp. 217–240, 2020.

[4] X. Li, "Mediating cross-cultural differences in research article rhetorical moves in academic translation: a pilot corpus-based study of abstracts," *Lingua*, vol. 238, Article ID 102795, 2020.

[5] F. Pan and B. Zheng, "Gender difference of hedging in interpreting for Chinese government press conferences: a corpus-based study," *Across Languages and Cultures*, vol. 18, no. 2, pp. 171–193, 2017.

[6] G. De Sutter and M. A. Lefer, "On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach," *Perspectives*, vol. 28, no. 1, pp. 1–23, 2020.

[7] M. Chen, Y. Zhang, M. Qiu, N. Guizani, and Y. Hao, "SPHA: smart personal health advisor based on deep analytics," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 164–169, 2018.

[8] M. H. Chen, S. T. Huang, J. S. Chang, and H. C. Liou, "Developing a corpus-based paraphrase tool to improve EFL learners' writing skills," *Computer Assisted Language Learning*, vol. 28, no. 1, pp. 22–40, 2015.

[9] T. Li and K. Hu, "Corpus-based translation studies and political discourse analysis," *Reappraising Self and Others*, vol. 1, pp. 13–51, 2021.

[10] C. Gu, "(Re)manufacturing consent in EnglishTarget," *International Journal of Translation Studies*, vol. 31, no. 3, pp. 465–499, 2019.

[11] M. Á. Candel Mora and O. Polyakova Nesterenko, "Building a corpus-based glossary of Spanish-Russian higher education for specialised translation," *Sendebar*, vol. 30, pp. 141–162, 2019.

[12] R. Xu, "Corpus-based terminological preparation for simultaneous interpreting," *Interpreting. International Journal of Research and Practice in Interpreting*, vol. 20, no. 1, pp. 33–62, 2018.

[13] N. Curry and A. Chambers, "Questions in English and French research articles in linguistics: a corpus-based contrastive analysis," *Corpus Pragmatics*, vol. 1, no. 4, pp. 327–350, 2017.

[14] T. Li and F. Pan, "Reshaping China's image: a corpus-based analysis of the English translation of Chinese political discourse," *Perspectives*, vol. 29, no. 3, pp. 354–370, 2021.

[15] H. Schendl, "3. Code-switching in Anglo-Saxon England: a corpus-based approach," *Multilingual Practices in Language History*, vol. 15, pp. 39–60, 2017.

[16] M. Pirhayati, "A parallel corpus-based study of collocations from English to Persian: criticism, and resolution," *Journal of Applied Linguistics and Language Research*, vol. 8, no. 3, pp. 109–126, 2021.

[17] M. V. Farahani and M. Sbetifard, "Metadiscourse features in English news writing among English native and Iranian writers: a comparative corpus-based inquiry," *Theory and Practice in Language Studies*, vol. 7, no. 12, p. 1249, 2017.

[18] D. Divjak, E. Dąbrowska, and A. Arppe, "Machine meets man: evaluating the psychological reality of corpus-based probabilistic models," *Cognitive Linguistics*, vol. 27, no. 1, pp. 1–33, 2016.

[19] V. Montalt, K. K. Zethsen, and W. Karwacka, "Medical translation in the 21st century-challenges and trends," *MonTI. Monografías de Traducción e Interpretación*, vol. 10, pp. 27–42, 2018.

[20] O. D. Kuzmina, A. D. Fominykh, and N. A. Abrosimova, "Problems of the English abbreviations in medical translation," *Procedia-Social and Behavioral Sciences*, vol. 199, pp. 548–554, 2015.

[21] M. Á. Jiménez-Crespo and M. Tercedor Sánchez, "Lexical variation, register and explicitation in medical translation: a comparable corpus study of medical terminology in US websites translated into Spanish," *Translation and Interpreting Studies*, vol. 12, no. 3, pp. 405–426, 2017.

[22] H. Ma, F. Yang, J. Ren et al., "ECCParaCorp: a cross-lingual parallel corpus towards cancer education, dissemination and application," *BMC Medical Informatics and Decision Making*, vol. 20, no. S3, pp. 122–212, 2020.