Application of Computer Intelligent Proofreading System in English Phrase Translation

Huang Jing(✉)

Wuhan Technology and Business University, Wuhan 430065, Hubei, China
20071208017@wtbu.edu.cn

Abstract. This paper studies the application of computer proofreading system in English word translation. The technology is closely related to the development of computer technology, structural theory, linguistics and other fields. With the improvement of computing power and explosive availability of multilingual information resources, from early dictionaries to regular dictionary translation combined with language knowledge, and then to statistical machine translation based on thesaurus. Machine translation technology has begun to come out of the ivory tower and provide real-time translation services for simple users. We notice that the English translation proofreading system of grammar and sentence patterns is an important problem of correct proofreading, and the intelligent proofreading system based on the improved word translation model can not solve the inconsistency problem. Through the establishment of semantic text and word combination conversion algorithm to simulate the syntactic part. Through semantic feature analysis and word combination, optimize the automatic translation algorithm, load the translation algorithm, and design the automatic translation system software based on the embedded environment. Therefore, cross compiler and multi-threaded word translation are used to design the translation system automatically. The development process of software module mainly includes system application, network translation, customer, translation organization and management, online translation interaction and other modular structures. Finally, the design system is verified. The verification results show that the system can meet the requirements of users for horizontal demarcation and improve the efficiency of demarcation.

Keywords: Intelligent proofreading system · Translation of English phrases · System development

1 Introduction

After entering the 21st century, with the rapid development, science and technology are becoming more and more frequent, and the amount of information is increasing sharply, which has greatly deepened people's understanding and demand for language, knowledge and intelligence. All words that may be misunderstood as "machine" as the

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2022, LNICST 465, pp. 161–171, 2023.
https://doi.org/10.1007/978-3-031-23950-2_18

subject has "intelligence" are problematic from the perspective of science and technology and pragmatics. "Machine intelligence", "machine learning", "Machine Cognition", "machine…" These words, as well as "man-machine symbiosis", "man-machine integration", "man-machine…" These specious words should have been avoided in the strict sense of science and technology. Although there are many "conventional" word usages in human society, it should only be "conventional" based on "agreement". If we understand and use these easily misunderstood words from their literal meaning, or even use them as the description of scientific research objects, I'm afraid they will not produce original research results and will only lead the research direction astray, which scholars should try to avoid. Using these words to name courses and even majors will only mislead people's children.

The so-called "artificial intelligence", human beings and their intelligence, should always be the first! Taking "artificial intelligence" as a technical means, human beings creatively discover, predict, recognize, summarize, abstract, calculate and construct all kinds of things in the social activities of understanding and transforming nature, so as to gradually and continuously improve the intelligence level of human beings, which is the correct direction of the progress of human society. No matter in scientific research, technological development, engineering practice, or in social communication, information exchange and daily life, excessively blindly believing in and relying on "machine intelligence" will inevitably lead to the overall degradation of the "human intelligence" of the dependent! In fact, such degradation phenomenon has been recognized and warned by some sober scholars in the world. Because it is beyond the purpose of this paper, it will not be repeated here.

In contradiction, the language barrier is becoming more and more serious. Relying solely on manual translation of massive data is not only inefficient, but also very expensive.

In essence, machine translation is the process of automatic transformation from one language to another. From the 1980s to 1990s, rule-based methods occupied the mainstream. Machine translation translated the source language through the grammatical rules designed by linguists, and achieved some results. Since the early 1990s, the acquisition of various bilingual corpora has become relatively easy. With the continuous improvement of computer processing ability, the implementation of statistical methods in technology and resources has become possible. During this period, various corpus based statistical machine translation methods have flourished, which greatly improves the translation system. Even in some specific fields, the quality of to a certain extent [1]. However, the current technologies and methods can not make the machine truly simulate the human translation process, let alone reach the level of human translation. However, the challenge and the demand of serving the public continue to inspire research institutions and researchers all over the world to devote themselves to the research and application of machine translation.

English is used to expressing complex concepts in long sentences, while Chinese is different. It often uses a number of short sentences for hierarchical narration. Therefore, when translating from English to Chinese, we should pay special attention to the differences between English and Chinese, and decompose the long sentences in English

into short sentences in Chinese. In the process of translating long English sentences, the following methods are summarized.

Firstly, when the content narrative level of English long sentences is basically consistent with that of Chinese, they can be translated into Chinese according to the hierarchical order of the original English, so as to make the translation basically consistent with the original English. Secondly, the expression order of some long sentences in English is different or even completely opposite to that in Chinese. At this time, translation must start from the back of the original text. In Chinese, attributive modifiers and adverbial modifiers often precede the modified ones.

In addition, there are mainly methods for English translation. In other words, when translating a long English sentence into Chinese, modifier phrases will be added before the elements of English are incorporated into the text in the usual Chinese order. But the decorative elements should not be too long. Otherwise, it will cause delay and contradiction [2].

Sometimes in English long sentences, the relationship between the subject or the subject sentence and the modifier is not very close. In translation, the clauses or phrases of long sentences can be transformed into sentences and described separately according to the habit of using short sentences in Chinese. In order to make the meaning coherent, it is sometimes necessary to add words appropriately, that is, to translate the whole English long sentence into several independent sentences by dividing the whole into parts, with the order basically unchanged and coherence.

The so-called word alignment refers to automatically extracting the corresponding relationship between words from bilingual parallel corpus. Figure 1 shows an example of word alignment of English and French sentence pairs. At present, most statistical machine translation systems use IBM model and hidden Markov model (HMM) for word alignment, but both models have a limitation on word alignment, that is, a source language word can only correspond to one target language word at most. Therefore, the word alignment generated by IBM model is usually also called asymmetric word alignment.

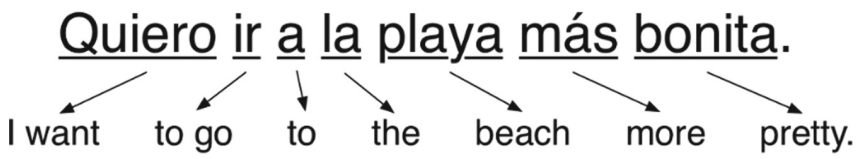

Fig. 1. Examples of word alignment in English and French sentence pairs

2 Related Work

2.1 Computer Intelligent Proofreading System

The advent of the big data era has brought the explosion of text information, and all kinds of traditional text analysis and processing work have begun to be replaced by

computers. The larger the amount of text data, the more errors it contains. It is particularly important to correct the errors in the text through proofreading. Traditional proofreading mainly relies on manual work, which can find and correct errors in the text. Manual proofreading has low efficiency, high intensity and long cycle. Obviously, it can not meet the needs of the rapid growth of the text. Intelligent proofreading system came into being under this background. With the development of machine learning and natural language processing technology, it is possible to use algorithm model to solve the problem of text proofreading. The research and development of intelligent proofreading system has greatly reduced the workload of proofreaders and made the previous heavy work mode simple, easy and efficient [3]. As shown in Fig. 2. Starting from the technical problems encountered in proofreading, this paper will lead readers to understand the technical methods of proofreading in the industry, as well as the technical principles and practical experience of cognitive intelligence laboratory in algorithm proofreading.

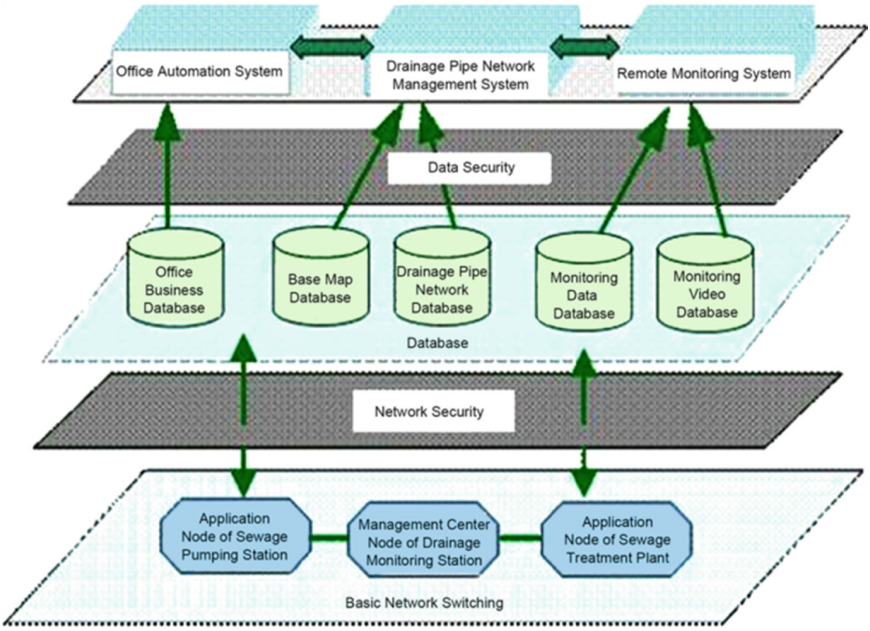

Fig. 2. Computer intelligent proofreading system

Automatic text proofreading is one of the main application fields of natural language processing, and it is also a research problem. The difficulties are mainly reflected in:

1. The distribution of true error samples is unknown. This problem is very different from other natural language processing problems or pattern recognition problems. Other natural language processing tasks have objective correspondence, that is, the model obtains the answer by identifying fixed patterns. However, in text proofreading, correct sentences and words are found from wrong sentences or words, and there is no objective correspondence. It can only be said that different proofreaders will get different answers,

and the final judgment of the answer is related to the cultural level and knowledge structure of proofreaders. Because it is "wrong" to find "right", a correct word may produce different wrong words due to different user input habits. In the error sample data set in the online real production environment, the distribution law from wrong words to right words will change at any time, which challenges the machine learning based on the independent distribution hypothesis.

2. A wide range of fields. Since the company's business serves all walks of life, the input text encountered in proofreading contains proper nouns in various fields, and a large number of professional knowledge dictionaries are needed to correct the input errors of input users from different industries [4]. In addition, in different professional fields, the character distribution of corpus varies greatly, so it is difficult to find the distribution balance of input training corpus when training the model, that is, how much corpus from different sources should be input into the model for training.

3. High performance requirements. It is embodied in recall rate, accuracy rate and model reasoning speed. Recall rate is the main evaluation basis for the performance of the proofreading system. It is used to describe how many errors in the real errors can be found by the algorithm. Accuracy is an important guarantee of good customer experience. Imagine a system with low accuracy and false positives throughout, which will greatly affect users' feelings and reduce users' efficiency. Fast model reasoning is an important requirement for the system to serve customers. If the speed of model proofreading is slow, the experience will be particularly poor for users who use proofreading services online.

2.2 Translation of English Phrases

Due to the differences in expression methods and habits between Chinese and English, translators are required to skillfully deal with the corresponding details and adjust the order of Chinese and English words in time. In general, when translating, the corresponding positions of Attributives and Adverbials must be adjusted to complete the translation.

For example, in Chinese, we usually use "she is cooking in the kitchen." To describe the action of the subject. However, in English, "she is cooking in the kitchen." To express the action of the subject [5]. The position of the adverbial of the two sentences has changed and transformed, which can meet the expression of the corresponding language and greatly improve the fluency of the sentence.

In English, if the position of the sentence attribute is a word, the word is usually placed before the modified word. In Chinese, this usage is the same. However, if postposition occurs in English, the translator needs to precede the attribute in the process of translation, so as to ensure the fluency of Chinese sentences and improve the quality and level of translation.

If it is necessary to translate more Chinese nouns into the front of the sentence, it can still be adjusted according to the English language. For example:

They have a lot of thing dificult to do.

他们有很多困难的问题需要解决。

In the process of translating English sentences, we should adjust the translation order according to the expression habits of Chinese, ensure the complete translation of

the meaning of the article according to the corresponding principles and the context of the article, and ensure the smoothness of the translated sentences, so as to improve the quality and level of English translation. In the process of adjusting and transforming the English word order, we should first translate according to the logical relationship of the sentence itself, and then adopt relevant translation skills, so as to meet the requirements of English translation.

Adjust according to the time sequence. In the translation between English and Chinese, the usage of the two is slightly different in terms of time order, so we need to understand this difference in translation, and we can translate according to the corresponding language habits to better meet the characteristics of the corresponding language. For example, in English, sentences describing time can usually be placed before or after the subject. However, when expressing in Chinese, it needs to be described according to the specific time and order of the action.

3 Application of Computer Intelligent Proofreading System in English Phrase Translation

At present, the research of statistical machine translation is developing in a diversified direction. A variety of knowledge sources and methods have been introduced into the research of statistical machine translation, such as language features, syntactic structure features, semantic features, word network, semantic network, etc., which has greatly promoted the research and development of statistical machine translation, and has basically met the application requirements in some specific fields. Basic realization: ordering + translation = decoder. Based on rules or statistical methods. In the middle process, there are a variety of smoothing and disambiguation methods, as shown in Fig. 3 below.

Fig. 3. Organize the sentence structure of the translation

An intelligent English automatic translation system platform is constructed, which is mainly composed of phrase translation combination based translation system, hierarchical phrase translation combination based translation system and syntax based translation system [6]. We mainly design the platform from two aspects: automation and modularization.

Optimization implementation: for phrase based statistical translation. They have certain advantages in ambiguity elimination, local sorting and decoding efficiency, reduce

the complexity faced by the machine translation system and show good model robustness. They are often used as the baseline of statistical machine translation system research.

Automation mainly means to minimize manual intervention. The operations from corpus preprocessing to model training, from test set input to translation result generation, from system initialization to automatic parameter adjustment are completely completed by computer; Modularization means that the translation system is an extremely complex and systematic process from front-end corpus processing to back-end translation result generation. At the same time, due to the different implementation and construction principles of multiple translation engines, there will be different requirements and representations in different operation stages. The technical route is shown in Fig. 4.

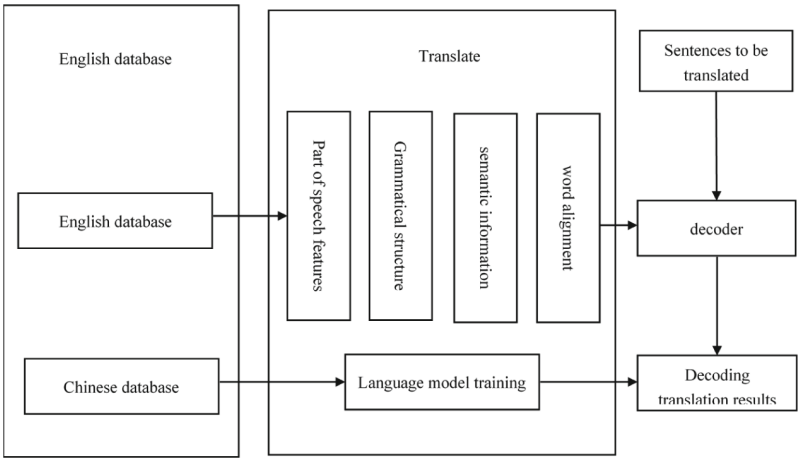

Fig. 4. Technical route of translation system

Section 5 below fig shows how one can effectively allocate limited attention resources when browsing images. The red area indicates that the visual system pays more attention to the target [7]. Of course, in the above cases, you can see the face, title and first sentence of the article.

In essence, the attention mechanism in deep learning is similar to the mechanism of human selective visual observation. The main purpose is to select important information from various sources for the purpose of this task.

The translation quality evaluation of machine translation system is an important factor to promote the rapid development of statistical, especially the development of automatic evaluation technology is the direct driving force of statistical machine translation methods. For the evaluation methods of translation quality, the current evaluation standards can be divided into two kinds [8]. One is subjective evaluation, that is, according to the given reference translation, the system translation results are scored manually, mainly including two indicators: Fluency and fidelity; Another method is automatic evaluation, that is to score the translation results according to a certain mathematical model. Similarly, the goal of automatic evaluation method is to achieve a balance between fluency and loyalty.

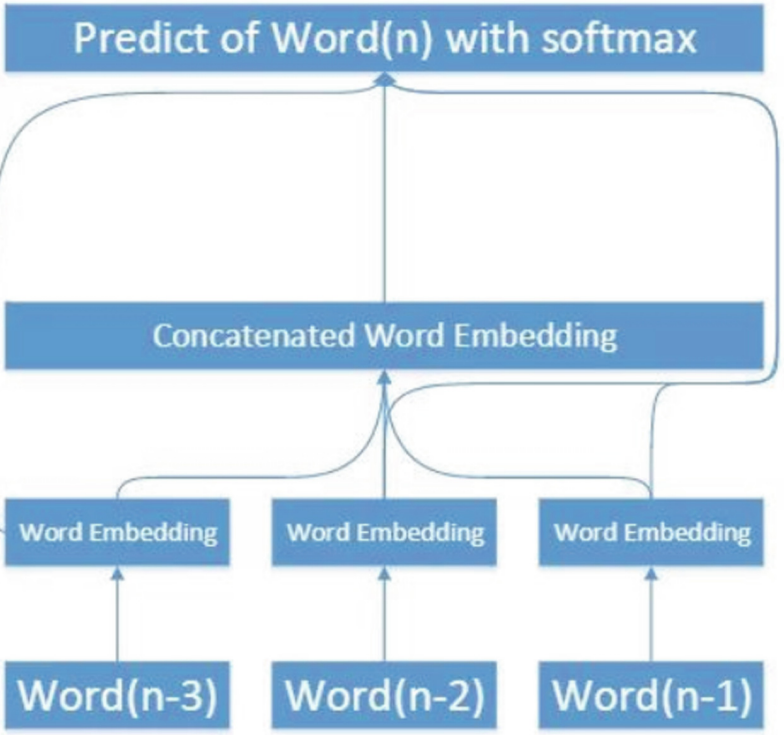

Fig. 5. Gets the score of each encoder's hidden state

At the stage of model reasoning, the mainstream autoregressive translation models are predicted by step. Each step requires the state of the previous, several or all artificial neurons. Every operation of transformer's decoder requires a lot of repeated calculations in the previous state.

Keep part of the calculated state, and make tradeoff in memory overhead and calculation overhead to realize acceleration.

Tensor2tensior gives the original open source implementation of Google's official transformer. Among them, the reasoning cache acceleration of vanilla transformer follows the above idea and obtains nearly twice the speed of long sentence reasoning. The calculation formula is shown in (1).

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{1}$$

where PN is the proportion of the number of n-grams appearing in the reference translation to the total number of n-grams in the candidate translation, $w_n = 1/N$, which means taking the geometric average of n-gram, and N is the maximum n-ary syntax

order (generally 4). BP is the length penalty factor to punish the short translation. The calculation formula is shown in (2).

$$BP = \begin{cases} 1 & c > r \\ e^{(1-r/c)}, & c \leq r \end{cases} \tag{2}$$

where, R is the length of the effective reference translation in the test corpus, which can be defined as three lengths: the length of the shortest reference translation, the length of the reference translation closest to the sentence length of the system candidate translation, the length of the longest reference translation, and C is the length of the system candidate translation [9].

Using the intermediate results generated by the attention mechanism of the translation model, a set of auxiliary pointer network is designed to realize the placeholder translation engine. After the key information is classified and replaced according to the part of speech, the translation obtained by the input model greatly reduces the missing translation and wrong translation rate of this kind of information.

With the continuous maturity of the technical framework, how to make more efficient use of resources has become a new challenge in academia and industry, including low resource learning, high concurrency or real-time model design, multilingual large-scale system establishment and so on. At the same time, the rapid iteration of machine translation technology has also triggered and driven more thinking and research, including the support of upstream models, the expansion of downstream applications, the integration of a priori common sense or domain knowledge, etc. a series of methodologies have continuously promoted the accelerated landing and improvement of artificial intelligence in natural language processing tasks.

4 Simulation Analysis

After the failure of rule-based system, new translation methods have been developed. They are not based on grammatical rules, but on probabilistic and statistical models.

The establishment of statistical translation system requires a lot of training data. Among them, the same text is translated into at least two languages. This kind of double translation text is called parallel material library [10]. Similarly, scientists in the 18th century found an Egyptian Hieroglyph from Greece on the 9th monument in losetta. The same pattern is also engraved on the stone tablets of Greece, ancient Egypt and folk characters. By comparing the contents of different language versions, modern archaeologists have the opportunity to explain the meaning and structure of Egyptian hieroglyphs that disappeared more than a thousand years ago. Today, it is an important milestone in the study of ancient Egyptian history. Computers can use parallel corpora to predict how texts are translated from one language to another.

Fortunately, there are many translators in the world. For example, the European Parliament has translated its agenda into 21 languages. Therefore, researchers will use these data to establish a translation system, as shown in Fig. 6 below.

The English automatic translation system is designed into two parts: algorithm and software. The design of the system software adopts the modules of dictionary information processing and dictionary program automatic control based on embedded environment.

English	Spanish
Resumption of the session	Reanudación del período de sesiones
I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.	Declaro reanudado el período de sesiones del Parlamento Europeo, interrumpido el viernes 17 de diciembre pasado, y reitero a Sus Señorías mi deseo de que hayan tenido unas buenas vacaciones.

Fig. 6. The translation system a large amount of training data

In order to extract the information that reflects the characteristics of the system information distribution rules, the type of information integration and intelligent scheduling is selected to realize the intelligent management and scheduling of the system. An automatic translation system is designed based on embedded ARM system, and the system network component interface is designed by TinyOS The information management system software is compiled in the Linux kernel, which improves the intelligent level of system management.

Establish your own language translation system. There is a presentation that can be translated between English and French. But not cowards and people with limited budgets. This is still a new technology with limited resources. Even with advanced computers equipped with modern video cards, it may take a month to learn language translation.

In addition, the rapid development of translation technology from sequence language to sequence language is difficult to achieve [11]. Many innovations (such as increasing attention and context monitoring) have significantly improved translation results, but they are innovative, not Wikipedia. If you want to switch from one sequence to another, you need to update it as technology advances.

The system network node interface based on TinyOS is optimized in application hierarchy, operation layer and data layer. In the design stage, the combined information is transmitted through the base station, the information is transmitted through the sensor node, and the high-speed a/d converter is converted into a vector for digital conversion. For this, the intelligent management function software of the automatic translation system is developed.

Current scope of application of machine translation: people who do not understand the source language want to know about the contents of some less professional information, such as e-mail, instant chat information, and some technical documents with high text repetition rate [12].

Idealized development after machine translation: through the continuous expansion and optimization of corpus, the development of search and capture technology and the expansion of the scope of statistical objects, in order to obtain the corresponding entry pairs with higher statistical probability and improve the accuracy of translation.

5 Conclusion

The core of the algorithm is to optimize the English automatic translation algorithm from the perspective of English translation combination. Cross compilation and program loading are used to load the algorithm into the information processing module of the system. On the basis of software development, the development of English automatic translation platform based on the combination of word and translation is promoted. Researching

and establishing an automatic English translation automation platform based on the combination of words and translation can further improve the intelligence and automation level of English translation. At present, machine translation methods are mainly statistical methods, and word based translation is the most mature of statistical methods, as the main component of statistical translation. In recent years, grammar based statistical machine translation has gradually become a hot topic in statistical machine translation research. The system design includes software design and machine translation algorithm By combining semantic feature analysis with word translation and optimizing automatic translation algorithm, the software design based on embedded automatic translation environment is realized.

References

1. Zhang, J., Luan, H., Sun, M., et al.: Neural machine translation with explicit phrase alignment. IEEE/ACM Trans. Audio Speech Lang. Process. **99**, 1 (2021)
2. Xl, A., Hy, B.: Parataxis or hypotaxis? Choices of taxis in Chinese–English translation. Lingua **251**, 103026 (2021)
3. Hidalgo-Ternero, C.M.: Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. MonTi Monografías de Traducción e Interpretación **32**, 154–177 (2021)
4. Ding, L., Wang, L., Liu, X., et al.: Progressive multi-granularity training for non-autoregressive translation **21**, (2021)
5. Min, J.: Research on the application of computer intelligent proofreading system in college English teaching. J. Phys: Conf. Ser. **1915**(3), 032078 (2021)
6. Guo, X., Zhu, Y., Zhang, J., et al.: Intelligent pointer meter interconnection solution for data collection in farmlands. Comput. Electron. Agricult. **182**(13), 105985 (2021)
7. Jamshidifarsani, H., Garbaya, S., Lim, T., et al.: Intelligent games for learning and the remediation of Dyslexia: using automaticity principles. IEEE Syst. Man Cybern. Mag. **7**(1), 15–24 (2021)
8. Gao, J., Guo, Z.: Application of text proofreading system based on artificial intelligence. In: Atiquzzaman, M., Yen, N., Zheng, Xu. (eds.) Big Data Analytics for Cyber-Physical System in Smart City: BDCPS 2020, 28-29 December 2020, Shanghai, China, pp. 722–727. Springer Singapore, Singapore (2021). https://doi.org/10.1007/978-981-33-4572-0_104
9. Zhu, W., Tan, C., Xu, Q., et al.: Heterogeneous identity trust management method based on risk assessment. J. Intell. Fuzzy Syst. **11**, 1–14 (2021)
10. Jansen, P.: Darmok and Jalad at Tanagra: a dataset and model for English-to-Tamarian translation (2021)
11. Lin, Z., Song, X., Guo, J., et al.: Peer feedback in translation training: a Quasi-experiment in an advanced Chinese–English translation course. Front. Psychol. **12**, 631898 (2021)
12. Peng, X.: Thinking path schema of English translation for Chinese classics: an empirical study on translation schema in translation courses. Engl. Lang. Teach. **14**(2), 56 (2021)