J. COMBIN. MATH. COMBIN. COMPUT. 127b (2025) 9653--9664

Journal of Combinatorial Mathematics and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on Translation System Based on cloud Computing Data Aggregation Algorithm

Xiaojing Dong[1], Li Yuan[2,✉]

[1] *Jilin Engineering Normal University, Changchun, Jilin, 130000, China*
[2] *Northeast Normal University, Changchun, Jilin, 130000, China*

ABSTRACT

Artificial intelligence technology has brought new breakthroughs to the field of machine translation. Through the introduction of cloud computing data aggregation algorithms, this paper proposes two translation methods, namely rules and corpus. At the same time, the translation system is studied with English as the research object. Based on the statistical translation method, the basic framework of the English translation system (ETS) is designed, including a preprocessing module, a source language matching module, a statistical decoding module, and a target translation generation module. And by introducing the k-means algorithm and the optimized k-means++ algorithm, ETS was studied. Combined with cloud computing technology, the ETS had a powerful data storage platform. Finally, a simulation experiment was carried out to test the performance of the system from three aspects: the average number and type of translation results, the success rate of translation in different languages, and the speed of online translation. First, the comparison method of the two algorithms was used to test them separately. The data showed that with the increase of vocabulary, the average number and types of translation results in the ETS have also increased. The system developed by k-means++ algorithm was 5.03 items higher than the average number of translation results of the system developed by k-means algorithm, and 1.93 items higher than the average number of categories. When testing the success rate of translation in six languages, the data showed that the average success rate of English translation in different languages remained at 94.34%. It was concluded that the success rate of using k-means++ was higher than that of k-means algorithm, and the k-means++ algorithm could make the translation system produce better results when running. Finally, the online translation speed of the common ETS and the ETS based on cloud computing technology were tested. The average online translation speed of the system under cloud computing technology was 40.46b/s under different translated text volumes, while the average online translation speed of the common system was 26.47b/s. It indicates that the efficiency of the ETS on the basis of cloud computing technology is high and the data processing capability is strong, which makes the system far more efficient than the ordinary translation system in operation and has obvious superiority.

✉ Corresponding author.
E-mail address: peteryuan2024@126.com (L. Yuan).

Received 05 March 2024; Revised 25 May 2024; Accepted 05 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-529

© 2025 The Author(s). Published by Combinatorial Press. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

Keywords: English Translation System, Artificial Intelligence, Online Translation Speed, Clustering Algorithms, Cloud Computing

1. Introduction

The deepening of economic globalization has led to an increase in the frequency and significance of linguistic and cultural interchange. In this particular setting, the translation system's effectiveness and efficiency play a crucial role as a vital tool for fostering communication. The quick growth of cloud computing in recent years has encouraged machine translation innovation. Data aggregation algorithms are the main research of the business environment; they are designed to enhance decision-making by extracting relevant information from large amounts of data. The data aggregation technique also has a wide range of potential applications in the translation system. The research object for this article is the English translation. English is a language used throughout the world and is significant in many different sectors. For instance, the computer sector uses English as the norm for some professional phrases. Computer English consequently emerged. Additionally, business English was developed in the area of international trade to improve communication and foster international cooperation. English language learners can now benefit greatly from intelligent electronic education systems that have evolved as a result of the rapid growth of Internet technology. Additionally, the development of ETS has encouraged international interactions as well as, to some extent, cross-cultural interactions.

English is an important common language for communication among countries today, and plays an essential role in the growth of world economy and culture. This field has also been attracting more and more attention and has been studied. Deng L designed a process-oriented assessment model on the basis of the preliminary findings of a project-based flipped learning model in a business English translation course. The collected data and results indicated that the process-oriented assessment model could greatly improve students' motivation and performance in the business English translation course [1]. Rivera-Trigueros I conducted a systematic literature review to identify the most commonly used MT systems currently in use. The results of the study showed that neural MT was the main paradigm in the current MT scenario and was the most commonly used system for Google Translate [2]. Liang J proposed a design method of a computer-aided translation system for English communication languages based on gray clustering evaluation. The findings showed that the designed system had the benefits of high online translation speed, pronunciation success rate and multilingual translation success rate [3]. These studies have certain reference value, but most of them stay in the theoretical aspect.

With the appearance of AI technology, intelligent translation has become mainstream today, and more and more people have used AI technology to study the field of ETSs. Gu S optimized the current computer-aided English translation teaching in view of the current situation and problems of professional English translation teaching in colleges. He designed a new computer-assisted translation system with reference to the cloud computing model to optimize English teaching. The experimental findings showed that the proposed new computer-assisted translation system not only achieved better translation performance, but also could effectively and quickly realize the intelligent translation of memory-assisted long character English [4]. Zhang B proposed a deep attention model and conducted experiments on Chinese-English, English-German and English-French translation tasks. Empirically, it was found that the deep attention model tended to produce more accurate attention weights as the attention layer was sufficiently increased. His in-depth analysis of translation of important contextual words further revealed that the deep attention model significantly improved the fidelity of systematic translation [5]. Song X constructed a model based on intelligent ETS on the basis of ML algorithm and improved multi-objective optimization algorithm to improve the effectiveness of English intelligent translation. His findings showed that the constructed translation model was highly intelligent and could meet the actual translation requirements [6]. These research results have certain reference significance for ETS, but they are not analyzed in combination with the current situation.

MT is now the choice of most people. Therefore, the development of intelligent ETS is now the focus of research in the field of English translation. This paper combined ML technology and cloud computing technology to study ETS, and introduced five basic methods of MT. On this basis, the statistical method was used to study ETS, and the source channel model was introduced. ML clustering algorithm was used to study the classification problem in ETS. Combined with cloud computing technology, the basic framework of ETS was designed. The data and results showed that the ETS based on k-means++ algorithm produced better translation effect, and the ETS based on cloud computing technology had stronger performance and higher operation efficiency than the ordinary ETS. Finally, a summary was made.

2. Intelligent ETS

2.1. MT Technology

Translation technology is a method used by translators to solve the problem of transmitting messages from the first language to be translated. Translation technology can tell the essence of the first language. It is applied to the level of words, phrases, clauses or sentences [7]. Before the emergence of AI technology, language translation was carried out manually, and some books and textbooks were carried out by professional translators. Due to the scarcity of professional translators, translation efficiency is slow in the period of manual translation. As the Internet era came, computer technology developed rapidly. The process of translating from one natural language to another through intelligent machines is called MT.

With the maturity of computer technology, MT technology continues to progress. After years of development, MT can generally be divided into two major translation methods: rule translation and corpus translation. Among them, rule-based methods can be divided into direct vocabulary translation, semantic-based translation, and intelligent transformation-based translation. The corpus-based translation methods can be divided into two types: statistical and example type. Corpus is the main requirement for the development of statistical MT and neural MT [8]. Figure 1 is about the method of MT.

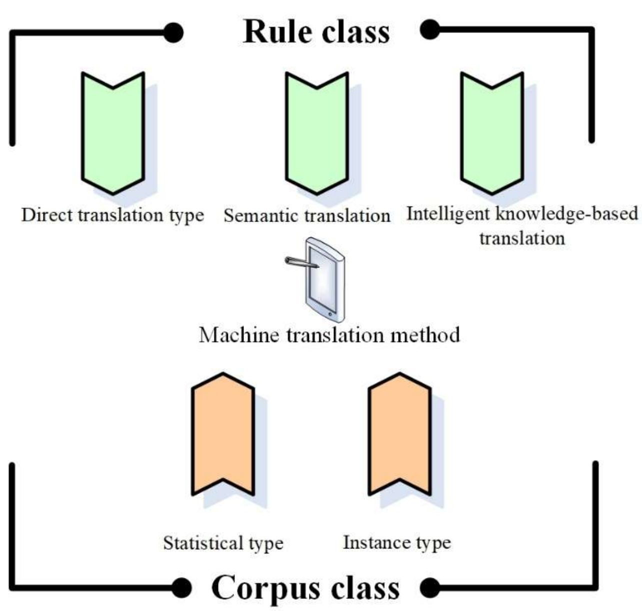

Fig. 1. Classification of MT methods

1) Direct vocabulary translation: Direct vocabulary translation refers to the comparison of the words and sentences to be translated through the data in the translation machine database, and they have the same and similar types. The source target is converted to the target language, and then the target language is output.

2) Translation based on semantic transformation: Semantic translation refers to dividing the source target to be translated into several relevant semantic elements, and finding the corresponding semantic representation by keyword matching and other methods. The MT system forms the overall target translation by testing the relationship between the semantic elements of each part and establishing its logical relationship.

3) Intelligent knowledge-based translation: Intelligent translation is the translation method developed from the former two. On this basis, the MT system is equipped with the common sense and professional knowledge that humans have, making the translation process more understandable. Larger knowledge base and dynamic selection technology are adopted to carry out semantic transformation and output multiple target translations. The intelligent knowledge-based translation method is the result of the growth of AI technology.

4) Statistical translation: Statistical translation is the most commonly used method of corpus-based translation, which refers to the construction of a knowledge base by dividing and tagging corpus. This knowledge base is transformed into data form for analysis. The words and sentences in the source target are probability calculated to form a certain form of data, so as to find the target sentence with the highest probability in the language database. The advantage of this method lies in its high cost performance and strong autonomy, which reduces manual intervention.

5) Case-based translation: Case-based translation refers to matching the most similar fragments in the corpus and combining each fragment into the best target translation through the existing empirical knowledge in the knowledge base and analogical translation.

2.2. ETS

With the fast growth of computer technology, the need for barrier-free communication between people is becoming more and more urgent. Therefore, it is extremely significant to build an ETS with high translation accuracy and quality [9]. English is the universal language in the world and the most widely used language in the world. Therefore, in order to promote exchanges between countries, the demand for talents in the English translation industry is increasing. This research on the translation system is mainly based on English translation.AI technology provides a new breakthrough for English translation. Through MT of English, multiple languages are converted, which is very difficult for human beings. The existing translation work has proved the potential of large-scale multilingual MT by training a single model that can translate between any pair of languages [10]. In translation, to understand the machine ETS, the concept of MT model must be first understood, and the target language can be generated through the machine language translation model. This paper introduces a statistical translation model. In statistical MT, relevant personnel put forward the concept of source channel model, which is the conversion between language model and translation model. Assuming the target language is α, the source language β is obtained through channel deformation. The goal of translation is to restore the source language β to the target language α. The process is shown in Figure 2.

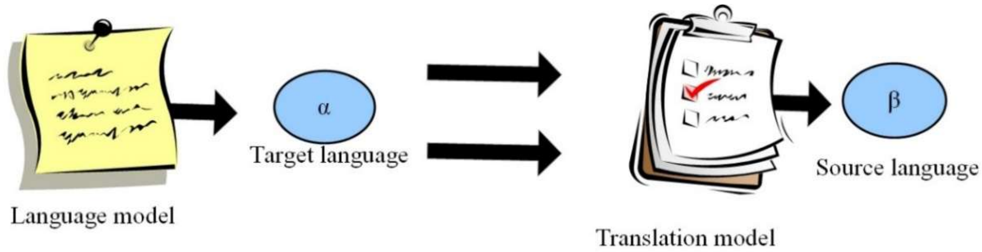

Fig. 2. MT channel model

The process of the translation model can be expressed by the following formula:

$$P(\alpha \mid \beta) = \frac{P(\alpha) \times P(\beta \mid \alpha)}{P(\beta)} \tag{1}$$

In Formula (1), the probability of $P(\alpha)$ target language is the language model. $P(\beta \mid \alpha)$ refers to the probability of translating the target language into the source language, which is called the translation model. $P(\alpha \mid \beta)$ is to calculate the probability of restoring the source language to the target language according to the probability model, that is, the translation process. This formula can deduce its decoding process, as follows:

$$\hat{\alpha} = \arg \mathrm{MAX}(P(\alpha) \times P(\beta \mid \alpha)) \tag{2}$$

In Formula (2), $\hat{\alpha}$ is the required target translation. By estimating the probability of $\hat{\alpha}$, the maximum value of $P(\alpha) \times P(\beta \mid \alpha)$ is found, which is the translation target. According to these two formulas, the core of the process of MT model is actually the process of probability estimation of the translation model. The efficiency of probability estimation of a system developed using different algorithms determines the performance of its system translation. Therefore, when developing an ETS, different algorithms have a certain impact on its system performance.

The process of MT includes various key activities, such as preprocessing, Part-Of-Speech marking, transliteration, word order alignment, translation, morphological analysis, etc. [11]. A complete ETS can be divided into the following modules: language preprocessing module, source language matching module, statistical decoding module, and target translation generation module. The preprocessing module refers to the grammar analysis, part of speech tagging and word statistics of the input text. The source language matching module refers to the segmentation through the matching model algorithm selection. In the corpus, the source language is matched. Statistical decoding module refers to the function of classifying and statistical translation after matching, calculating the maximum probability of translation model, and finally performing database statistics. The target translation generation module is integrated and sorted after module translation to select the best combination, and the final translation is output. The framework structure about the ETS is shown in Figure 3.

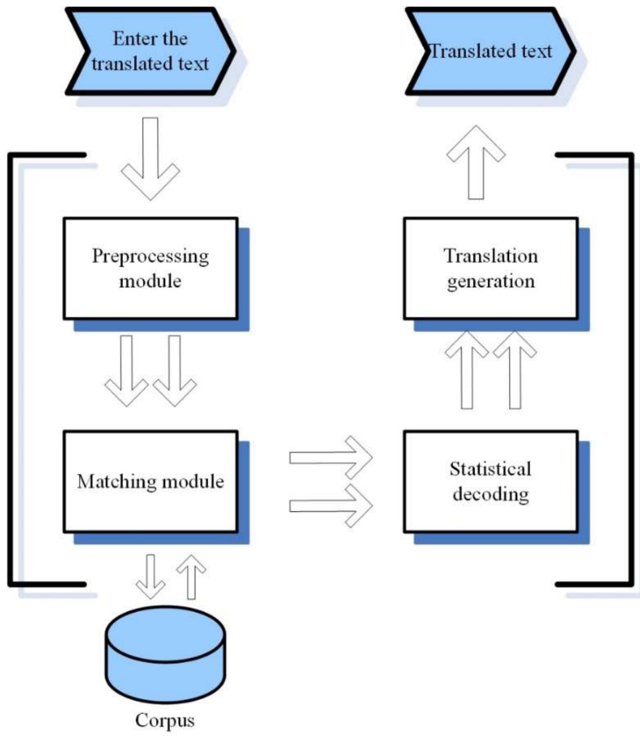

Fig. 3. Basic framework of ETS

Figure 3 shows the basic framework design of the ETS. The ETS divides the input source language into more accurate phrases and sentences, and matches them from the corpus knowledge base. The translation model is used to match, and the algorithm is used to count the sentences or words in the target database that have the highest probability of matching with the source language, so as to generate the target translation.

2.3. ML

AI is the science that studies the laws of human intelligent activity and is the evolution of computer systems, which are capable of performing tasks that normally require human intelligence [12]. ML is the branch that belongs to the AI technology. The performance of ETSs depends on the technology used to develop the system as well as on the algorithms. Therefore, this paper proposes a clustering algorithm in ML to solve this problem.

2.3.1. Clustering Algorithm. ML refers to imitating and learning human behavior through machines. Clustering algorithm is one of ML, which is divided into supervised algorithm and unsupervised algorithm. The clustering algorithm belongs to unsupervised learning. The biggest difference between supervised learning and unsupervised learning is whether there are training samples for learning. The former has samples and the latter does not. In the unsupervised algorithm, the clustering algorithm refers to dividing the given data into different categories. By calculating the similarity between samples, the maximum similarity is divided into one category, which is applied to the ETS. When the source language is matched with the corpus, the same and similar sentence phrases are classified into one category. The most similar category is selected for translation. The commonly used clustering algorithms include k-means algorithm, k-means++ algorithm, and so on. Figure 4 shows the iteration diagram of clustering algorithm.

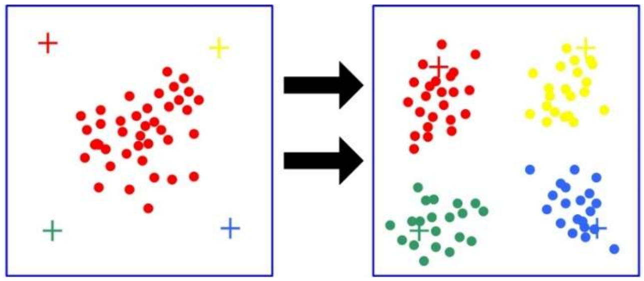

Fig. 4. Iterative diagram of clustering algorithm

From Figure 4, the running process of the clustering algorithm can be learned. This paper introduces the clustering k-means and k-means++ algorithms. The basic principle of k-means algorithm is to define k categories in a space model composed of N nodes. Each category has a central node for initialization. The node nearest to the center is classified by distance calculation. The iterative method is used to continuously update the values of each cluster center until the adjacent two times do not change. The distance calculation formula is as follows:

$$D = \mathrm{MIN}\sum_{i=1}^{N}\left\|x_i - y_j\right\|^2 (j = 1, 2, \cdots, k) \tag{3}$$

Formula (3) is derived from the Euclidean distance formula. Among them, y_j is the initialized data center and x_i is the data point. The k-means algorithm also has its shortcomings. It not only needs to determine the value of k in advance, but also is sensitive to the initial centroid and abnormal data. Therefore, the k-means++ algorithm is introduced here to deal with the shortcomings of the k-means algorithm. The principle and process of k-means++ algorithm are similar to that of k-means algorithm. Among them, the main optimization is the selection of initial centroid points, and the k-means++ algorithm randomly selects a point as the cluster center. When the cluster center is less than k, the shortest distance D from each point to the center is calculated. The shorter the distance is, the more likely it is to be selected as the next cluster center. Then, the next cluster center is selected. The method used here is the roulette method. The formula is as follows:

$$P = \frac{D(x)^2}{\sum_{x \in N} D(x)^2} \tag{4}$$

In Formula (4), P is the probability of the next cluster center. One measure of the cluster effect is the cluster effect SSE (Sum of Squared Error), which indicates the sum of squares of the post-cluster clusters from the cluster center. Table 1 shows the table of calculated values of SSE regarding k-means algorithm and k-means++ algorithm.

Table 1. SSE value calculation of k-means algorithm and k-means++ algorithm

Serial number	K-means (SSE)	K-means++(SSE)
1	1433	119
2	784	123
3	366	121
4	2339	125
5	987	120

It can be learned from Table 1 that the SSE value of the k-means algorithm changes greatly each time and is unstable. However, the SSE value of k-means++ algorithm changes little each time, and is basically stable at about 120. It can be learned that the SEE value of k-means++ algorithm is relatively stable. Therefore, it can be concluded that the clustering effect of k-means++ algorithm is better than that of k-means algorithm. This paper selects the clustering algorithm to develop the ETS.

When matching the source language, the clustering algorithm is used to classify the similar word meanings.

2.4. Cloud Computing

ETS needs not only good word meaning classification technology, but also powerful data storage and processing functions. Aiming at this problem, this article proposes to use cloud computing technology to develop an ETS. Internet-based cloud computing has the most powerful computing architecture [13]. The main advantages of cloud computing are its powerful data computing capability, large scale servers, high reliability and scalability, and low cost. The key technologies of cloud computing include virtualization technology, distributed computing and database technology, resource management and data storage technology. Cloud computing has a certain architecture. When using cloud computing technology to develop ETS, according to the architecture of cloud computing, the ETS architecture is designed. First of all, the front desk should have a user interface module, an English translation service module, and a resource management module. The background system should need deployment and matching tools, server management, etc. Figure 5 shows the relationship between cloud computing and ETS.

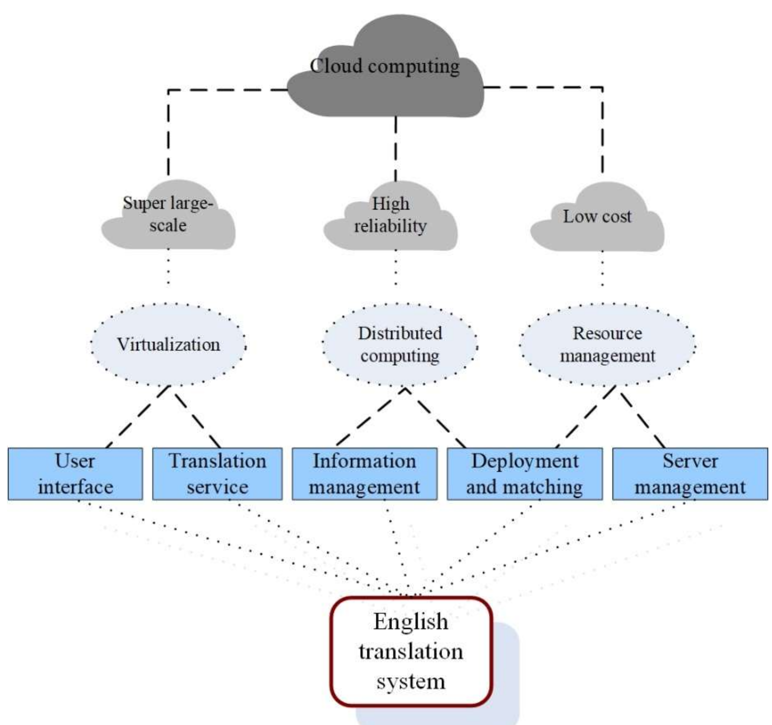

Fig. 5. The relationship between cloud computing technology and ETS

The use of cloud computing system can effectively deploy dynamic information, and it also has resource monitoring function. The following is a simulation experiment of the ETS.

3. Simulation Experiment of ETS

The importance of MT systems is increasing due to the massive production of online texts in different disciplines and the inability of traditional translation methods to meet the translation needs around the world [14]. In this paper, ML clustering algorithm was used to study the ETS, and cloud computing technology was used to calculate the probability of restoring the source language to the target language according to the probability model in the statistical translation method mentioned above. According to the input text, it was matched with the corpus. The clustering algorithm has strong classification ability. Cloud computing has powerful data processing, storage and computing capabilities. Therefore, this paper tested the ETS simulated by the k-means algorithm and the improved k-means++ algorithm, and tested the system performance from the average number and type of translation results, the success rate of different languages, and the online translation speed. First of all, the ETS simulated by k-means algorithm and k-means++ algorithm was tested by inputting different numbers of words from the system. The average number and type of translation result entries were calculated, as shown in Figure 6.

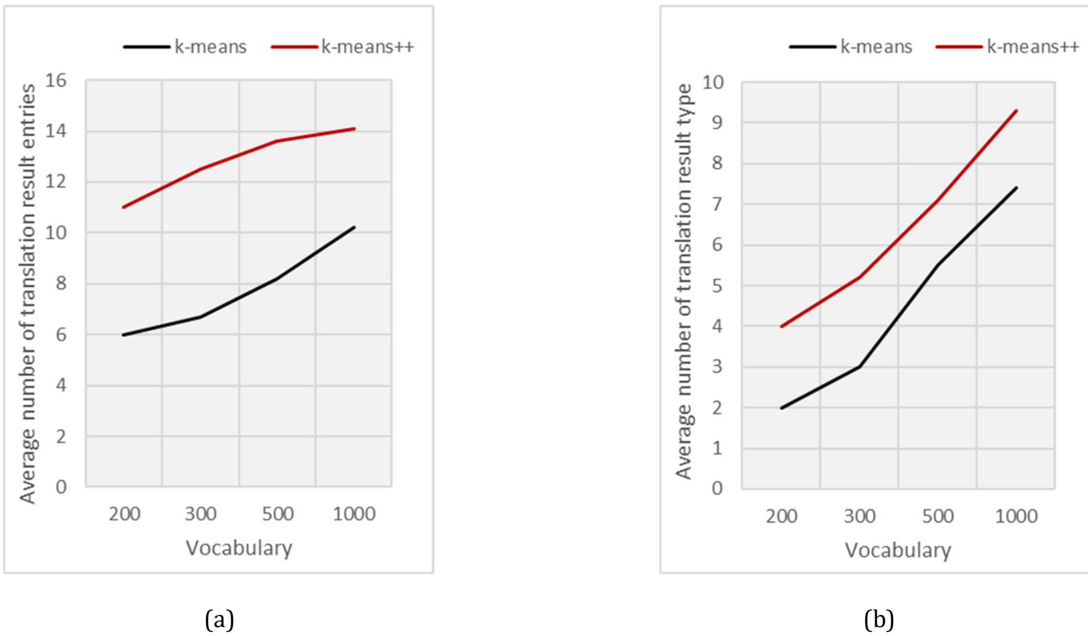

Fig. 6. Comparison of translation results between k-means algorithm and k-means++ algorithm system

Figure 6(a) Average number of translation results of different words

Figure 6(b) Average translation results of different words

From the data in Figure 6, the k-means++ algorithm had a higher translation system result value than the k-means algorithm, regardless of the average number or type of translation results. The data in Figure 6(a) showed that with the increase of vocabulary, the average number of translation result entries also increased. The system developed by k-means++ algorithm had an average number of translation results 5.03 higher than the system developed by k-means algorithm. According to the data in Figure 6(b), with the increase of vocabulary, the average number of translation results was also increasing. The types of English translation results developed by k-means++ algorithm in the figure were also higher than those of k-means algorithm, with an average difference of 1.93. Therefore, it can be obtained that the ETS developed based on k-means++ algorithm performed higher translation accuracy than k-means algorithm. Figure 7 is the calculation of the success rate of English translation in different languages.

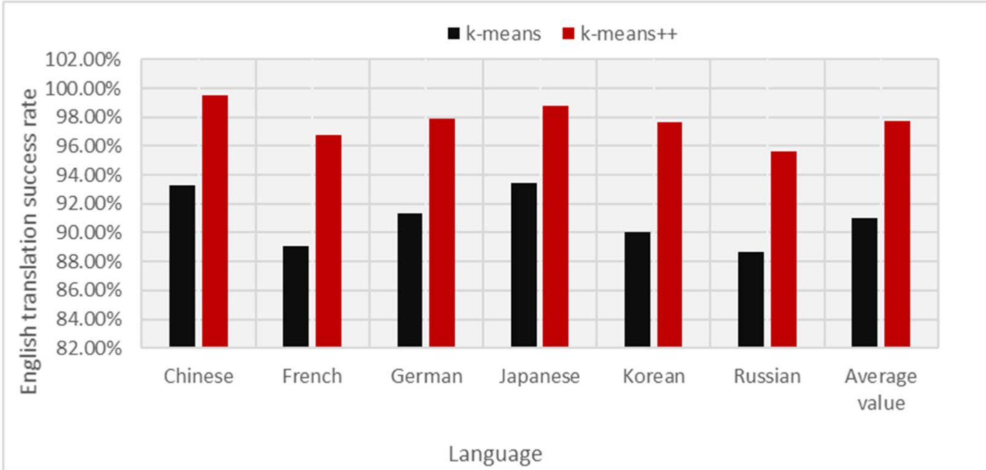

Fig. 7. Matching success rate of English translation in different languages

Figure 7 shows the success rate of English translation in Chinese, French, German, Japanese, Korean and Russian. Two algorithms are used for translation, and the findings are presented in the figure. The average success rate of English translation in different languages remained at 94.34%. The data in the figure showed that the average successful matching rate of k-means++ algorithm translation was 97.72%, while that of k-means algorithm translation was 90.97%. The results showed that the ETS of clustering algorithm had high performance, and the translation effect of k-means++ algorithm was better than that of k-means algorithm. Therefore, the ETS selected k-means++ algorithm when matching words, so that the translation system could produce better results when running. From the perspective of online translation speed, the online translation speed of the common ETS and the online translation speed of the ETS developed using cloud computing technology were simulated and analyzed. The online translation speed under different translated text volumes was calculated, as shown in Figure 8.

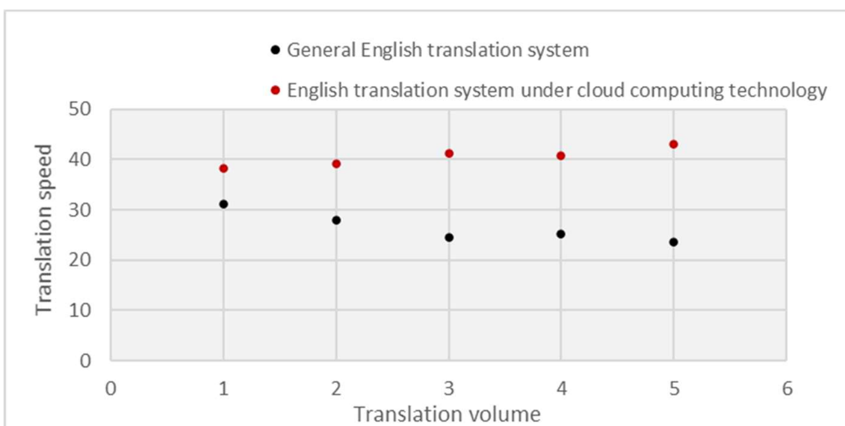

Fig. 8. Online translation speed of different ETSs

Figure 8 is about the online translation speed calculation of the translation system under the general ETS and cloud computing technology at different translated text volumes. 1, 2, 3, 4 and 5 in the figure corresponded to 100, 200, 300, 400 and 500 respectively. The data in the figure showed that the online translation speed of the ETS based on cloud computing technology was generally higher than that of the ordinary ETS, and the translation volume was 100-500. The online translation speed of cloud computing technology ETS and common ETS was 38.13b/s, 39.14b/s, 41.23b/s, 40.67b/s, 43.12b/s and 31.16b/s, 27.87b/s, 24.56b/s, 25.09b/s and 23.67b/s respectively. By calculating the average of the two, the average online translation speed of the system under cloud computing technology was 40.46b/s, while the average online translation speed of the ordinary system was 26.47b/s, which was 13.99b/s lower than the former. Therefore, it can be learned that the ETS based on cloud computing technology has high efficiency and strong data processing ability, which makes the efficiency of the system in operation far higher than that of the ordinary translation system, and has obvious advantages.

4. Conclusions

The problem of MT is an open research area. The establishment of an independent, effective and efficient translation system is a challenge [15]. This paper started with MT, and introduced the concept of MT and some methods of MT. The statistical translation method based on corpus was adopted to study ETS. The process of MT model was proposed. Its core was the process view of probability estimation of translation model, and its basic framework was designed. The ETS should include language preprocessing module, source language matching module, statistical decoding module and target translation generation module. Subsequently, ML clustering algorithm was introduced, and k-means and k-means++ were introduced to classify English translation matching. It was proposed to use cloud computing technology to provide a powerful operation platform for ETS for data processing and storage. Finally, simulation experiments were conducted to test three aspects: the average number and type of translation result entries, the success rate of translation in different languages, and the speed of online translation. The overall data showed that the ETS based on k-means++ algorithm had higher performance than the ETS based on k-means algorithm, which involved a wider range of language data categories, and had a higher success rate in different languages. Finally, it was concluded that the ETS based on cloud computing had higher performance and data computing and storage capacity than the ordinary ETS.

Funding

A study on the path of college English Curriculum ideology and politics. Doctoral research start-up fund project of Jilin Engineering Normal University (Project number: BSSK202301).

References

[1] Deng L. The Process-Oriented Assessment Model of Business English Translation Course in a Flipped Learning Context[J]. Higher Education Studies, 2020, 10(4): 1-11.

[2] Rivera-Trigueros I. Machine translation systems and quality assessment: a systematic review[J]. Language Resources and Evaluation, 2022, 56(2): 593-619.

[3] Liang J, Ma P. Design of computer aided translation system for English communication language based on grey clustering evaluation[J]. Journal of Computational Methods in Sciences and Engineering, 2021, 21(6): 2023-2033.

[4] Gu S, Li X. Optimization of computer-aided English translation teaching based on network teaching platform[J]. Computer-Aided Design and Applications, 2021, 19(S1): 151-160.

[5] Zhang B, Xiong D, Su J. Neural machine translation with deep attention[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 42(1): 154-163.

[6] Song X. Intelligent English translation system based on evolutionary multi-objective optimization algorithm[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(4): 6327-6337.

[7] Simanjuntak M B, Lumingkewas M S, Sutrisno S. ANALYSIS OF TANGIANG ALE AMANAMI (OUR FATHER) USING THE TECHNIQUES OF TRANSLATION[J]. JOURNAL OF ADVANCED ENGLISH STUDIES, 2021, 4(2): 70-75.

[8] Deep K, Kumar A, Goyal V. Development of Punjabi-English (PunEng) Parallel Corpus for Machine Translation System[J]. International Journal of Engineering & Technology, 2018, 7(2): 690-693.

[9] Yuan Z, Jin C, Chen Z. Research on language analysis of English translation system based on fuzzy algorithm[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(4): 6039-6047.

[10] Fan A, Bhosale S, Schwenk H . Beyond english-centric multilingual machine translation[J]. The Journal of Machine Learning Research, 2021, 22(1): 4839-4886.

[11] Dubey P. The Hindi to Dogri machine translation system: grammatical perspective[J]. International Journal of Information Technology, 2019, 11(1): 171-182.

[12] Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics[J]. Genome medicine, 2019, 11(1): 1-12.

[13] Srivastava P, Khan R. A review paper on cloud computing[J]. International Journal of Advanced Research in Computer Science and Software Engineering, 2018, 8(6): 17-20.

[14] Omar A, Gomaa Y. The machine translation of literature: Implications for translation pedagogy[J]. International Journal of Emerging Technologies in Learning (IJET), 2020, 15(11): 228-235.

[15] Kumar A, Goyal V. Hindi to Punjabi machine translation system based on statistical approach[J]. Journal of Statistics and Management Systems, 2018, 21(4): 547-552.