Hindawi
Mathematical Problems in Engineering
Volume 2023, Article ID 9810964, 1 page
https://doi.org/10.1155/2023/9810964

Retraction

Retracted: Cross-Context Accurate English Translation Method Based on the Machine Learning Model

Mathematical Problems in Engineering

Received 19 September 2023; Accepted 19 September 2023; Published 20 September 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Q. Zhang, "Cross-Context Accurate English Translation Method Based on the Machine Learning Model," *Mathematical Problems in Engineering*, vol. 2022, Article ID 9396650, 11 pages, 2022.

Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 9396650, 11 pages
https://doi.org/10.1155/2022/9396650

Research Article

Cross-Context Accurate English Translation Method Based on the Machine Learning Model

Qiang Zhang

Department of Foreign Language Studies, Anyang University, Anyang, Henan 455000, China

Correspondence should be addressed to Qiang Zhang; 001330@ayxy.edu.cn

Received 26 July 2022; Revised 24 September 2022; Accepted 27 September 2022; Published 17 October 2022

Academic Editor: Zaoli Yang

The era of big data and cloud computing has come, communication between different languages is becoming more and more common, and the barriers between languages are becoming more and more prominent. As the most important means to overcome language barriers, machine translation will play an increasingly important role in modern society. The previous machine translation technology has more or less disadvantages. The accuracy of translation is too low, which is a huge bottleneck hindering the further development of machine translation technology. Therefore, based on this, we can consider modeling the cross-context accurate English translation model based on the machine translation model and rely on the working principle of machine learning. This experiment shows that the translation accuracy of our method reaches 94.2%, which is higher than 39.5% of the benchmark method. This shows that the method in this paper can reduce the influence of other factors, ensure the accuracy of cross-context English translation to a certain extent, and meet the performance improvement requirements of the English translation system.

1. Introduction

With the popularization of computers, the rapid development of computer application technology and the deepening of global integration, communication barriers between groups using different languages have become increasingly prominent. Aiming at this problem, machine translation is a new subject, and it is also a hot research field of artificial intelligence. Machine translation involves many fields, such as mathematics, linguistics, and computer science. It is a typical interdisciplinary subject [1, 2]. It is no exaggeration to say that after entering the 21st century, almost everyone who lives in the information network era has to deal with machine translation directly or indirectly. No matter in science and technology, business, or politics, machine translation is undoubtedly a very important practical subject [3, 4]. The traditional machine translation method uses pipeline successive operations to mark the part of speech and analyze the syntax of the original corpus, so as to obtain the syntax structure of English language, which leads to the iterative transmission of errors between translation tasks and the reduction of the accuracy of structured examples, resulting in the reduction of the accuracy of English language and literature translation [5]. At present, English noun phrase recognition methods include the machine learning-based recognition method, statistics-based recognition method, and rules-based recognition method. The rule-based recognition method is mainly obtained automatically from corpus or compiled by experts, which has the advantage of easy understanding, but it is poor in generality, time-consuming, and easy to produce ambiguity. The method based on statistics transforms the problem of noun phrase recognition into the problem of labeling similar words. This method is simple and flexible and does not depend on the specific language model. At present, it is a popular mainstream translation algorithm. However, this method is based on a large number of sample data, which is prone to fitting problems. With the rise of artificial intelligence and machine learning methods, noun phrase recognition based on neural network is applied to English noun phrase recognition. However, due to the complex rules of English grammar, the accuracy of English noun phrase recognition needs to be improved.

Cross-context accurate English translation is widely used: (1) For professional translators, it can reduce their postediting time; (2) for the end users, they can know the quality of the translation output by the machine translation system; and (3) for machine translation systems, it can be used as one of the bases for ranking candidate translations [6]. Regarding the estimation of machine translation at the sentence level, the related research generally regards it as a supervised regression problem, and the main research work focuses on feature extraction and feature selection [7]. Feature extraction refers to extracting some grammatical and semantic features related to quality estimation from source sentences and machine translation sentences and may also use some external resources, such as alignment tools and language models. Feature selection refers to the selection of the feature subset which has the best prediction effect on the sentence quality of machine translation from all the extracted features mentioned above. The commonly used feature selection algorithms include Gaussian process, heuristic, and so on [8]. With the development of machine learning, some researches apply the algorithms in machine learning to the process of feature extraction and then input the extracted features into statistical machine learning models alone or together with other traditional features, such as support vector regression and linear regression. At present, the commonly used machine translation models are of two types [9, 10]. In English translation, the quality of the translated text is an important criterion to measure the translation result, which is mainly reflected by the characters, spelling errors, inconsistent expressions, and lexical and grammatical errors of the translated text [11]. At present, there are various types of machine translation tools, but their translation accuracy is low. When proofreading English translation results, too much attention is paid to the accuracy of phrases and syntax, but the proofreading of contextual coherence is directly ignored. However, the current research mainly focuses on improving the translation accuracy, making the translated products lightweight, and making a good experience for users, but it lacks some attention to the quality control of cross-contextual English translation [12]. Therefore, we can consider modeling the cross-context accurate English translation model based on the machine translation model. Based on this, this paper makes a detailed analysis and discussion of cross-context accurate English translation methods, focusing on cross-context accurate English translation based on the machine learning model. The purpose of this paper is to provide some valuable references for improving the quality and accuracy of English translation and achieving the goal of cross-context accurate English translation.

This paper explores the cross-context accurate English translation method based on the machine learning model. The innovations of this paper are as follows:

(1) Innovation in topic selection, combining the machine learning model with "cross-context accurate English translation method" to enrich the traditional English translation theory system. It also provides a new idea for the development of machine learning and has a certain reference value for professionals in the field of English translation.

(2) It breaks through the traditional translation concept. From examples, machine learning and analogy machine learning can make computer programs search for previously involved problems and simulate people's thinking ability to solve such problems before. Using the optimization algorithm based on gradient descent, the extraction time of feature engineering is reduced, and the accuracy of English translation in cross-context is improved.

Starting from the overview of machine learning and aiming at the application of machine learning methods in cross-context accurate English translation, this paper makes an in-depth analysis of cross-context accurate English translation methods based on the machine learning model, which is structured as follows. The first section is the introduction. This part mainly expounds the research background and significance of cross-context accurate English translation method based on the machine learning model and puts forward the research purpose, method, and innovation of this paper. Section 2 is a summary of the related literature of machine translation, summarizing its advantages and disadvantages, and putting forward the research ideas of this paper. Section 3 is the method part, focusing on the cross-context accurate English translation method based on the machine learning model. Section 4 is the experimental analysis. In this part, the experimental verification is carried out on the data set to analyze the accuracy of the modeled cross-context English translation.

2. Related Work

The era of big data and cloud computing has come, communication between different languages is becoming more and more common, and the barriers between languages are becoming more and more prominent. As the most important means to overcome language barriers, machine translation will play an increasingly important role in modern society. Yuval, Matton, and David believe that natural language, as the main bearer of information, and how to effectively solve the language barrier between different languages has become an important issue that cannot be ignored in human society. Machine translation is an effective method to solve this problem by using computer to realize automatic switching between multiple languages. Although the quality of machine translation is not perfect at present, machine translation is useful, and it has been used in many aspects such as distribution, browsing, communication, and information acquisition since its birth. The communication between different languages is becoming more and more common, and the barriers between languages are becoming more and more prominent. As the most important means to overcome language barriers, machine translation will play an increasingly important role in modern society [13]. Deborah, Beatrice, and others proposed a corpus-based method. With the help of the computer's automatic processing of real example sentences

in the corpus, the limited rules can be extended to an infinite degree, which can avoid the need to manually write translation rules. In the corpus-based method, the instance-based machine translation system directly uses the translation of similar instances in the corpus as a template, and after necessary corrections, it generates the final translation, making the full use of the resources of the corpus to avoid obscure sentences [14]. Ni et al. proposed that in today's retrieval technology, it is more advisable to use the similarity between sentences to measure the degree of difference between two entities. Sentence similarity can be divided into three levels: grammar, semantics, and pragmatics. The grammatical similarity is to perform grammatical analysis on two sentences, establish two grammatical trees, and grammatically calculate the similarity between the two sentences; however, it only considers grammar but not semantics. The defect is destined that it can only be used as a reference data. Semantic similarity refers to the degree of overlap between two sentences in the superficial sense, and its function is higher than grammatical similarity but lower than pragmatic similarity [15]. Harat et al. analyzed that people have recognized the limitations of machine translation and no longer expect machine translation to completely replace human translation. Just using the advantages of fast speed and large processing volume of machine translation, as an aid to human translation, can greatly improve the work efficiency of translators [16]. Beatriz and Helena improved the traditional rule-based machine translation model, using the English machine translation model based on semantic network, in the specific implementation process, using the phrase synthesis semantic statistical English machine translation method based on vector mixture [17]. Quang-Phuoc N et al. believed that the channel model of machine translation can be understood as follows: suppose someone wants to speak a sentence in the target language, but the sentence in the source language is spoken, which is an encoding process and statistical machine translation is to deduce this sentence from this sentence in the source language, it is a decoding process [18]. Poncelas et al. believed that in the statistical-based method, the acquisition of translation knowledge is completed before translation, and the translation process relies less on the corpus, while in the instance-based method, only the source language sentences are used before translation. Doing simple preprocessing does not obtain deep grammar and syntactic structure information and continues to obtain relevant knowledge from the corpus in the process of translation [19].

Judging from the representative research literature listed above, the current research literature is concise and comprehensive, and often on the basis of appropriate analysis, the core viewpoints and main conclusions and measures are put forward clearly. However, the above literature also has the characteristics of being too theoretical, and there is little research on how to carry out accurate English translation across contexts. Therefore, it is necessary to analyze and research on machine learning models, and there is also possibility of further development of such type of research in the future. Based on this, this paper conducts a detailed analysis and discussion on the cross-context accurate English translation based on the machine learning model, focusing on the method of cross-context accurate English translation under the influence of the machine learning model. It aims to provide some valuable references for improving the quality of cross-context accurate English translation and realizing the purpose of cross-context accurate English translation.

3. Methodology

3.1. Machine Learning Related Theory. When studying the theory of machine translation of natural language processing, we began to focus on the related machine learning methods. This is because the object of machine translation is natural language, and the process of human cognition of language has not been clearly studied. Therefore, there is still a long way to go to achieve ideal and high-quality machine translation. Machine translation is one of the scientific and technological problems to be solved in the century. The main difficulty is the ambiguity of natural language at all levels. It is difficult to fundamentally break through the ambiguity problem, which will involve the difficulty of processing and the speed of translation. For these closed-type errors, a limited puzzle set is often defined in advance, linguistic features (commonly used linguistic features include adjacent words, parts of speech, and dependent syntax tree) are extracted, and the text is converted into numerical representation. Then, based on these features, the classifier is trained by machine learning algorithm, and once the training is completed, it can be used to detect and correct the text [20].

Learning is to use all the attribute values of the known instance set as the training set of the learning algorithm, deduce a classification mechanism, and then use this classification mechanism to judge the attributes of a new instance. Training input data and testing input data are collectively called input, and input X and expected output T can be obtained through our research object. The training sample set composed of the training input data and the corresponding output of the system is given to the learning machine for model training. Learning machine is equivalent to finding a certain dependency between input X and expected output T, and the purpose of finding this relationship is not only to show good performance to the trained samples but also to be able to adapt to "new samples" as well. Its ability to adapt to samples is called generalization performance. The basic system flow chart of machine learning is shown in Figure 1:

The main purpose of machine learning research is to use computers to simulate human learning activities. It is to study methods for computers to identify existing knowledge, acquire new knowledge, continuously improve performance, and realize their own perfection. Learning here means learning from data, and there are three categories of supervised learning, unsupervised learning, and semi-supervised learning [21]. Supervised learning generally includes classification and regression, unsupervised learning generally includes probability density estimation, clustering, and dimensionality reduction, and semisupervised learning commonly used algorithms EM and constrain.

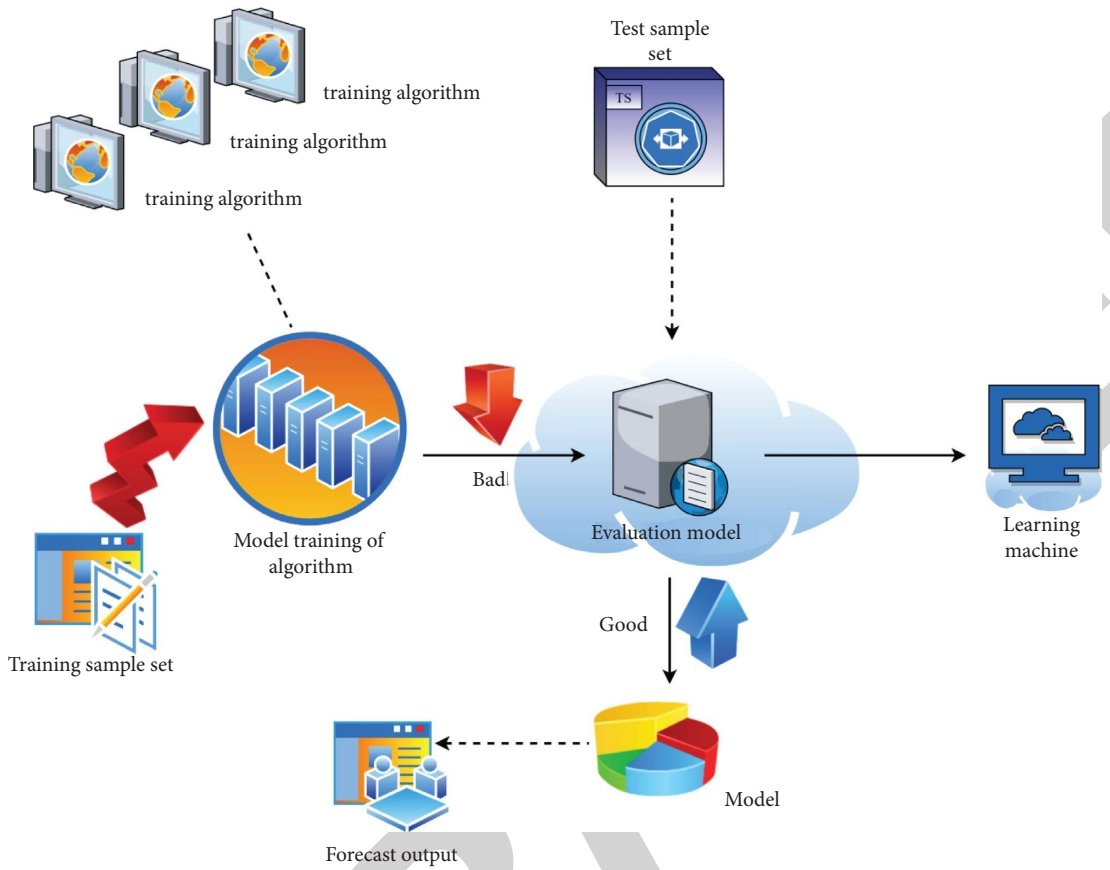

Figure 1: Basic system flow chart of machine learning.

3.2. Application of Machine Learning Methods in Cross-Context Accurate English Translation. When decoding, each step of the decoder may be far away from some words in the sentence in the source language; for example, if the first word on the target is the translation of the first word in the sentence in the source language, then in the first step of decoding, the distance from the first word at the source is equal to the length of the source sentence. If the source sentence is very long, this is obviously not conducive to the accurate translation of the first word and will greatly affect the translation of subsequent words. In fact, when translating sentences in the target language, the correlation between the words to be translated and different words in the source sentence is different, and this correlation will change with each step of decoding. If this encoding method is used to encode a fixed-length vector, the correlation between each step and all input words is unchanged during decoding, which is contrary to the actual translation process. The corpus used in the intelligent translation model plays an important role [22]. The corpus can be used to store bilingual phrase data, accurately label the parts of speech of short words, standardize the function of each phrase, and improve the timeliness and accuracy of the automatic phrase recognition algorithm in the English machine translation process. Figure 2 shows the information flow of the phrase corpus.

The alignment model in the machine translation model adopts hard alignment, and each target word will correspond to zero one or more words in the source sentence. This alignment model is trained separately, and after the training is completed, it will be used together with other components in the machine translation model to translate the target language sentence. In the field of machine translation, the intelligent identification of phrases is the key technology, which can meet the requirements of the tuning of translation samples and the accurate alignment of parallel corpora. Using the intelligent identification of phrases can effectively reduce grammatical ambiguity. When studying the theory of machine translation of natural language processing, we began to focus on the related machine learning methods. This is because the object of machine translation is natural language, and structural ambiguity is a difficult point in the current English translation field, which needs to be solved by the part-of-speech recognition algorithm. To achieve ideal and high-quality machine translation, it is necessary to fundamentally break through the difficulty of processing and the speed of translation. The complexity of translation is based on such difficulties, and we have to seek solutions to the problems in various ways. From examples, machine learning and analogy machine learning can make computer programs search for previously involved problems and simulate people's thinking ability to solve such problems before, which is undoubtedly a good method. Learning this

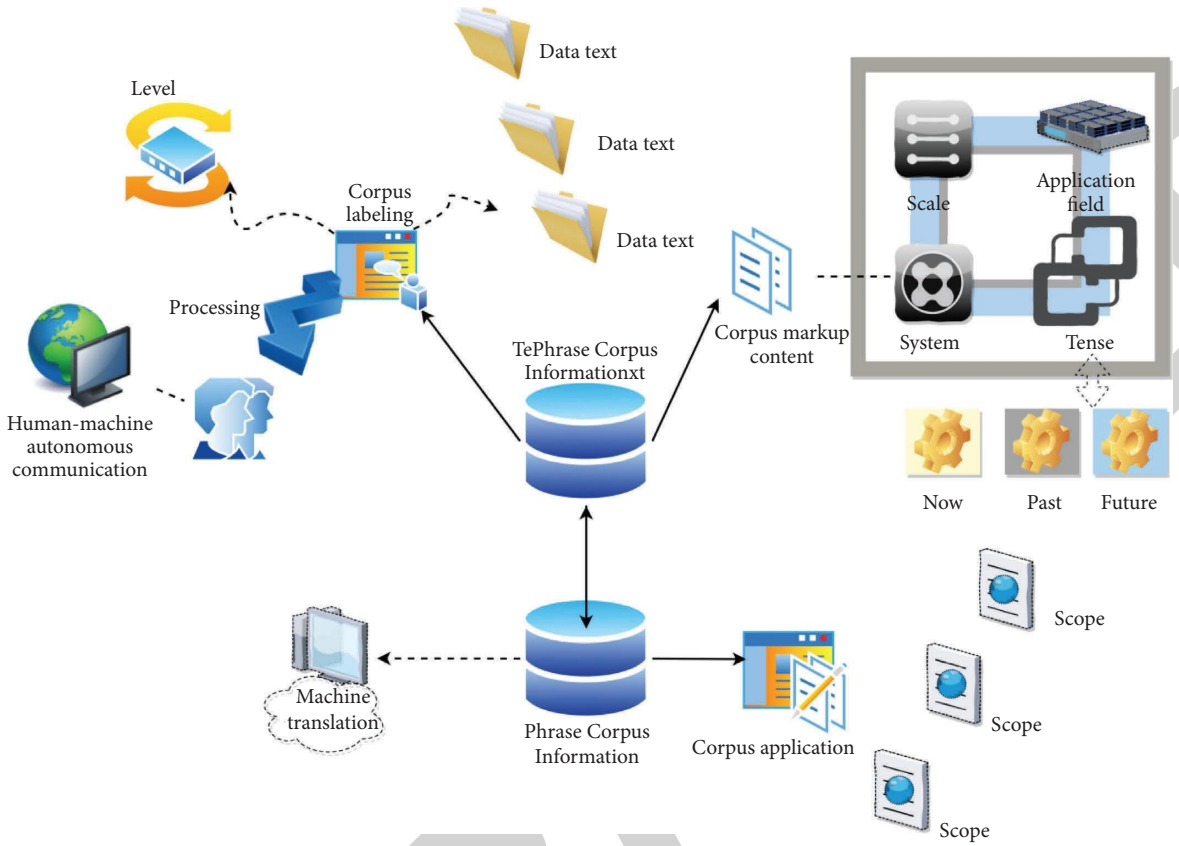

Figure 2: Phrase corpus information flow.

form of learning provides the computer program with a set of positive examples and negative examples of a certain concept, and the computer program induces a general concept description, making it suitable for all positive examples and excluding all negative examples. Analogy learning is a common learning method by comparing similar things. For example, when we encounter a new problem to deal with, we always recall the similar problems that have been dealt with in the past and find a solution that is closest to the current situation. At present, applying this idea to machine learning has got analogy learning.

3.3. Construction of the Translation System Based on the Machine Learning Model. Classification is a fundamental task in data mining. The traditional method is to find classification rules by means of mathematical statistics, pattern recognition, and other methods. However, in many cases, the prediction accuracy of traditional methods is greatly affected by the quality of training data and the limitation of professional knowledge, especially for those prediction problems whose essential laws are not fully understood, and the effect is not very satisfactory. Machine learning techniques are a common approach to solving these problems. Machine learning is to use all the attribute values of the known instance set as the training set of the learning algorithm, deduce a classification mechanism, and then use this classification mechanism to judge the attributes of a new instance. The commonly used classification methods include Bayesian classification and decision tree classification. However, a common problem of these methods is that their classification accuracy is not high when the amount of data is large. For this important problem, the algorithm provides us with an ideal solution. The implementation of the Adaboost algorithm relies on changing the data distribution, adjusting the weight of each sample according to whether the classification of each sample in each training set is correct or not and also the accuracy of the previous classification. After modifying the weights of the new data set, it is sent to the following classifier, and then trained, and finally the classifiers obtained from each training are fused to obtain a strong classifier. Using an algorithm, some less important features of the training data can be excluded and key features are focused. Its algorithm flow chart can be represented as shown in Figure 3:

Among machine learning algorithms, K-means is a classic algorithm. According to the research in this paper, we can apply it to sample segmentation. The specific operations are as follows. First, determine the initial clustering center c as

$$c = \left\{o_1^{(1)}, o_2^{(1)}, o_3^{(1)}, \ldots o_c^{(1)}\right\}, \quad (1)$$

where $o_i^{(k)}$ represents the i cluster center in the k iteration. According to the similarity with various centers, divide the samples into the nearest category, calculate the average of all samples in each category, and regard it as a new cluster center as

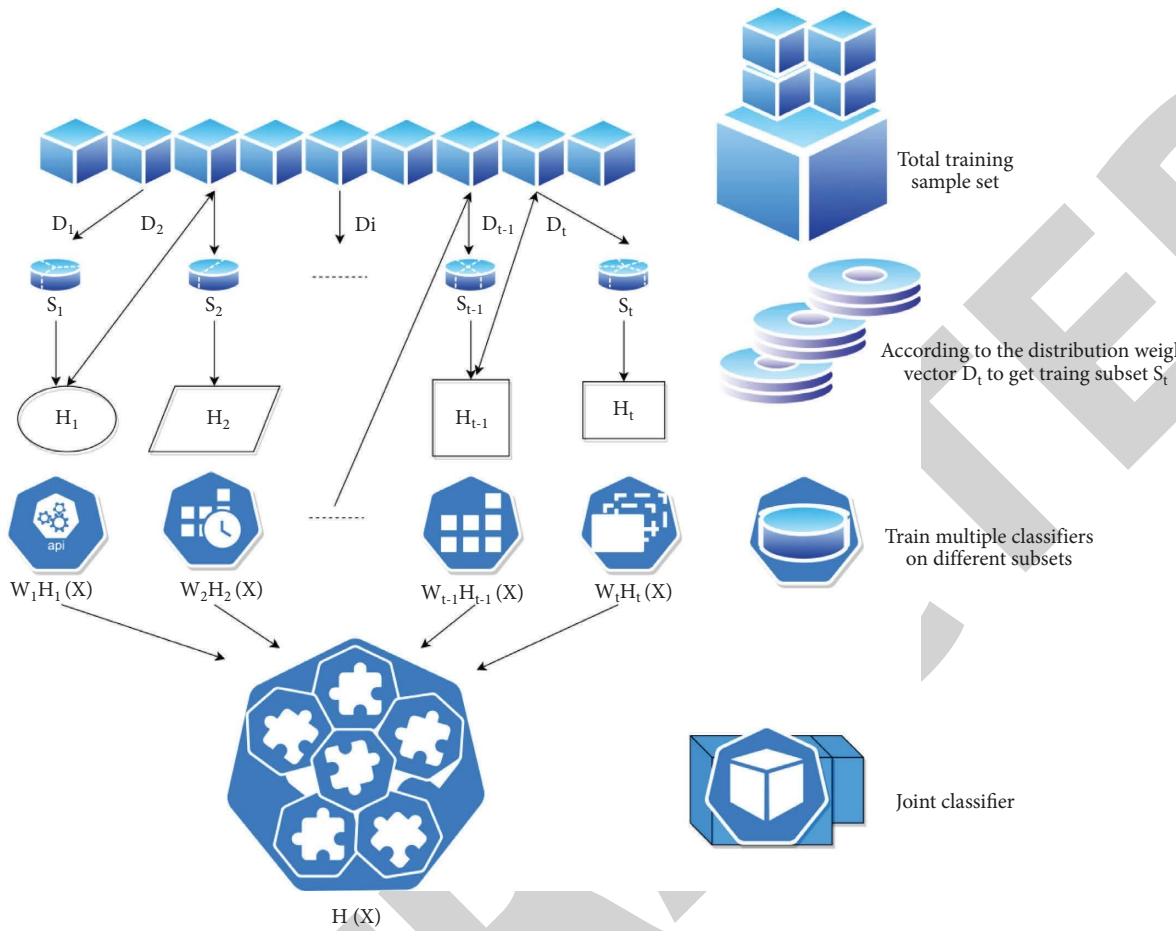

Figure 3: AdaBoost algorithm flow.

$$o_j^k = \frac{1}{n} \sum_{x_j \in s_j^{(k)}} x_j, \quad j = 1, 2, 3.... \tag{2}$$

The final value of the iteration stays at

$$o_j^{(k+1)} = o_j^{(k)}. \tag{3}$$

Otherwise, $k = k + 1$ proceeds with

$$o_j^k = \frac{1}{n} \sum_{x_j \in s_j^{(k)}} x_j - x_{k+1}^2 - x_k, \quad j = 1, 2, 3.... \tag{4}$$

For the input X of the system and the corresponding expected output T, N distributed observation samples are obtained: (x_1, t_1), (x_2, t_2), ..., (x_n, t_n). The prediction function of each input by machine learning is $f(X, \eta)$, so the expression of expected risk is as follows:

$$R(\eta) = \int L(t, f(x, \eta)) dF(x, t). \tag{5}$$

In which $L(t, f(x, \eta))$ is the loss function value of the prediction function f(X, η) of the actual expected output T with respect to X under the appropriate generalized parameter η, and $F(x, t)$ represents the joint probability distribution of input and output. It can be seen that the expected risk represents the accumulation of losses between all the actual sample values and the predicted sample sets of the system we study, and it represents the prediction performance of the prediction function for all the sample sets. In order to optimize the expected function, it is necessary to minimize the expected risk, which not only has a good effect on the trained samples but also has a better generalization performance. For a single specific sample, it indicates the difference between the predicted value of the model and the real sample value. The smaller the loss function, the more accurate the model is to predict the sample. Risk minimization is to minimize the average loss function of all sample points in the training set. The higher the empirical risk, the better the fitting degree of the model $f(x)$ to the training set. The smaller the empirical risk is, the more complex the model decision function is, and the more parameters it contains. When the empirical risk function is small to a certain extent, the phenomenon of overfitting occurs. It can also be understood that the complexity of the model decision function is a necessary condition for overfitting, which fully describes the consistency between the existing system model samples and the predicted results after training, and its expression is

$$R_{erm}(\eta) = \frac{1}{n} \sum_{i=1}^{n} L(t_i, f(x_i, \eta)). \tag{6}$$

We know that even if a very good model is established in a limited set of known samples, it cannot guarantee its good performance in "new samples." Therefore, the generalized loss bound is introduced into statistical learning, and the formula is as follows:

$$R(\eta) \le R_{erm}(\eta) + \psi(h/n). \tag{7}$$

To minimize the expected risk, that is, to minimize the sum of experience risk and confidence risk, structural risk minimization is put forward, and the formula is expressed as

$$\min_f R_{srm}(\eta) = \min_f R_{erm}(\eta) + \lambda\psi\left(\frac{h}{n}\right). \tag{8}$$

The posterior probability of the sample to the category is estimated by the conditional probability and prior probability of the category, so as to realize the judgment of the category to which the sample belongs. According to Bayes theorem, we get

$$P\left(c_i|d_j\right) = \frac{p\left(d_j|c_i\right)P\left(c_i\right)}{P\left(d_j\right)}. \tag{9}$$

It is the same for all categories. The prior probability can be obtained by simple estimation, usually taking the ratio of the number of c_i samples to the number of samples in the whole training set. Using the independent hypothesis, calculate $P(d_j|c_i)$ as

$$P\left(d_j|c_i\right) = \prod_{s=1}^{t} P\left(w_{sj}|c_i\right). \tag{10}$$

We regard the approximation as having zero error mean for these n samples, namely:

$$\sum_{j=1}^{K} \left\| t_j - o_j \right\| = 0. \tag{11}$$

In order to avoid too little contribution to the experimental results caused by small attribute values in the training process, the gait feature matrix is normalized to $[-1, 1]$, and the normalization function expression is

$$y_{ij} = \frac{\left(y_{\max} - y_{\min}\right)\left(x_{ij} - x_{\min}\right)}{\left(x_{\max} - x_{\min}\right)} + y_{\min}. \tag{12}$$

Use the RELM algorithm to train the data. The prediction results are calculated. The average absolute error and root mean square error are used to evaluate the accuracy and effectiveness of the prediction model, and the expression is

$$RMSE = \sqrt{\frac{1}{S}\sum_{i=1}^{s}\left(y_i - t_i\right)^2},$$
$$MAE = \frac{1}{s}\sum_{i=1}^{s}\left|y_i - t_i\right|, \tag{13}$$

where s is the number of training samples, y_i is the predicted output value, and t_i is the expected output value. Therefore, the regression fitting model is established, and the main steps are as follows: (1) forming a sample set, activating the function, and setting an appropriate number of hidden layer nodes; (2) divide D into training sample set and testing sample set, and randomly set input weight matrix and offset vector; (3) obtaining the output matrix of the hidden layer of the neural network; (4) by calculating the optimal solution of the output weight of the model; (5) substituting the obtained parameter values into formula (7) to obtain the predicted output; and (6) evaluate the analytical prediction model by calculating the results of RMSE and MAE.

The range of similarity is $[0, 1]$, and the semantic similarity between different words W_1 and W_2 is

$$\mathrm{Sim}_{\mathrm{semantic}}(W_1, W_2) = \max_{i=1,2,\ldots,n, j=1,2,\ldots m} \mathrm{Sim}\left(S_{1i}, S_{2i}\right). \tag{14}$$

The semantic similarity of two words is the highest value of the similarity of concepts between two words. The conceptual similarity of words is described, and the similarity between the operation sememes p_1 and p_2 is adopted as

$$\mathrm{Sim}\left(p_1, p_2\right) = \frac{\alpha}{d + \alpha}. \tag{15}$$

For the number linear model, the judgment model of multifeature thinking is adopted. For a given sentence, a translation is formed, and its maximum entropy translation model is

$$e_I^J = \sum_{m=1}^{M} \lambda_m h_m\left(e_I^J, f_I^J\right). \tag{16}$$

The logarithmic linear model has strong expansibility, can set corresponding features according to different target requirements, and can apply various linguistic methods to machine translation.

4. Result Analysis and Discussion

In order to verify the quality of cross-context accurate English translation based on the machine learning model, this paper adopts support vector regression and multilayer perceptron model and makes six experiments on two different QE data sets. The difference of different experiments lies in the difference of input features and models. Experiment CBOWE is selected as the training word vector in this experiment. When training, the dimension of word vector is set to 2048, the window size is set to 10, the number of negative samples used is 10, and the number of iterations is set to 15. The training set used to train the word vectors of the words at the source and target ends, respectively, adopts the source and target language materials in the parallel language materials of the training machine translation model. The language materials include WMT16, Europarl v7, and QE task1 corpus, and there are about five million sentence pairs in total. The direction of QE data set used is Chinese to English, and SVR is adopted in the model. Pearson correlation coefficient (de-en, SVR) is shown in Table 1.

Each row represents a class of input features, and the first column is the meaning of the features. Pearson's correlation coefficient ranges from −1 to 1. The larger the value, the

Table 1: Pearson correlation coefficient (de-en, SVR).

Features	Dev	Test
Embedding average	0.452	0.439
17 baseline features	0.501	0.488
Embedding + baseline	0.526	0.518

Table 2: Bilingual data set and WMT16 QE data set.

Corpus	Quantity
Parallel double statement pair	5000000
QE training set	25000
QE validation set	1000
QE test set	2000

Table 3: "Predictor-estimator" network parameter settings.

Parameter	Predictor	Estimator
Number of hidden layer nodes	512	128
Word vector dimension	512	512
Batch size	128	128
Optimization function	Lazyadam	Lazyadam
Source vocabulary size	140000	140000
Target vocabulary size	140000	140000
Node type	LSTM	LSTM

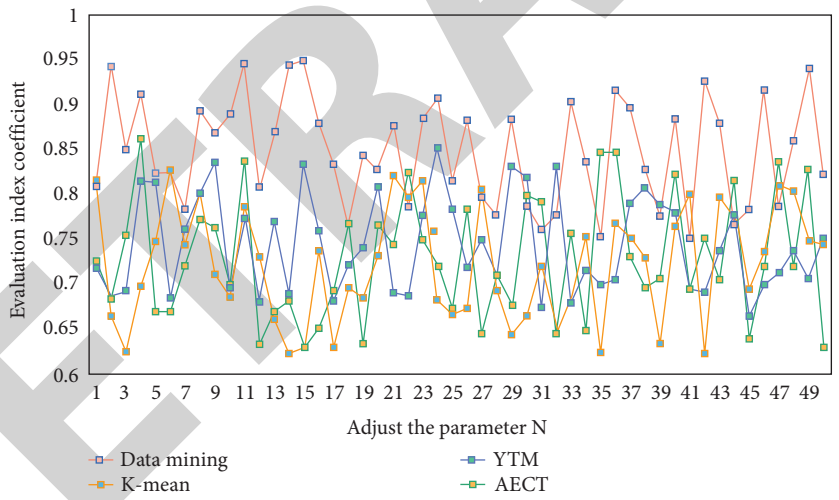

Figure 4: K-mean validity.

more accurate the predicted HTER is. Table 2 gives the relevant quantitative information of the experimental data.

The predictor module and the translation module set different model parameters, and their parameter settings are shown in Table 3.

In this experiment, the parameter n in the K-mean algorithm is adjusted so that $N = 1$, and the correct rate and normalized mutual information value of the corresponding K-mean result are obtained as shown in Figure 4, and the estimated number of clusters is shown in Figure 5.

It can be seen from Figures 4 and 5 that when $a = 0.095$, the K-mean algorithm proposed in this paper can not only obtain the correct number of clusters but also maximize the clustering effectiveness. This is because when the value of a is too large or too small, the local information of the data set will be hidden, resulting in the detector failing to drift to the peak near the probability density function, and the clustering result will become worse. First, it is assumed that each cluster in the sample space obeys some known probability distribution rule. Then different probability density functions are used to fit the statistical histogram in the samples. Continuously move the position of the center (mean) of the density function until the best fitting effect is obtained. The peak point of these probability density functions is the center of clustering. Then, according to the distance between each sample and each center, the category to which the nearest cluster center belongs is selected as the category of the sample. Because there are relatively few "BAD" tags in the training corpus, the model tends to predict the result

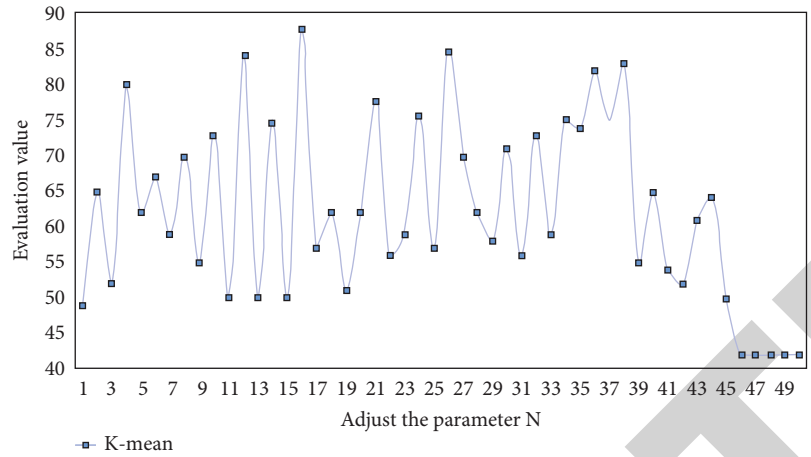

Figure 5: Estimates of the number of clusters.

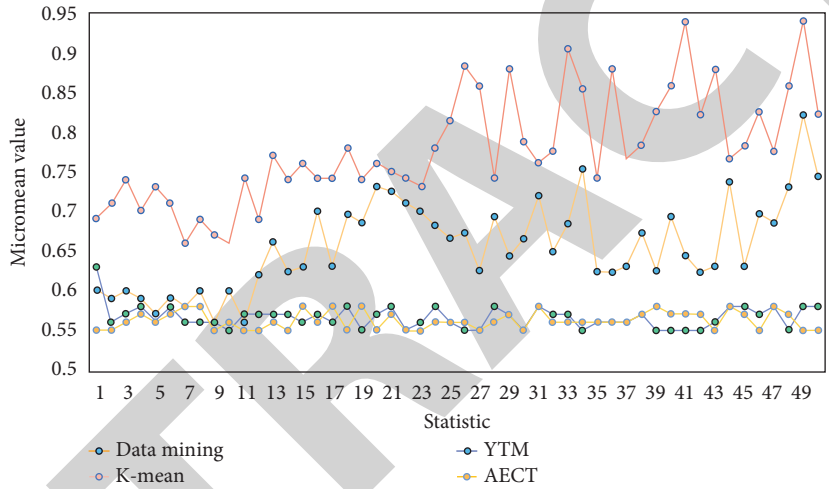

Figure 6: Comparison of the micro averages of the four algorithms.

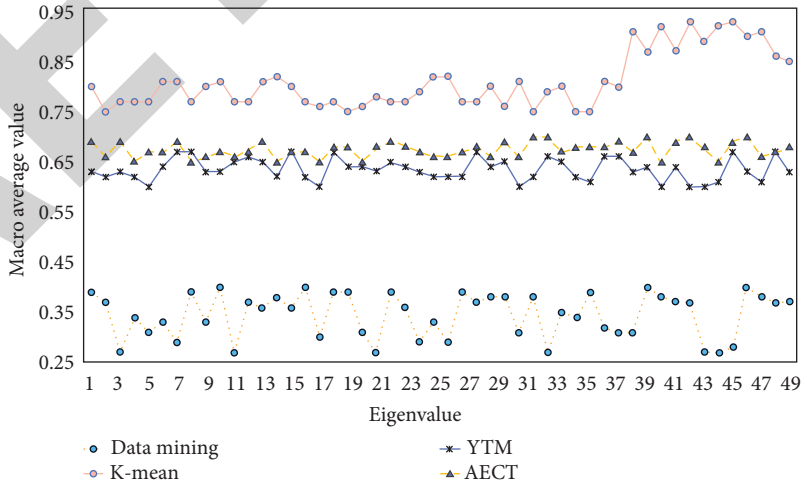

Figure 7: Comparison of macro averages under different algorithms.

as "OK," so the whole performance of the system is improved by increasing its F1 value through experiments. The experimental performance comparison is shown in Figure 6.

It can be concluded that the feature improvement method based on machine learning can obtain a micro average of 94.5%. In order to better reflect the advantages of

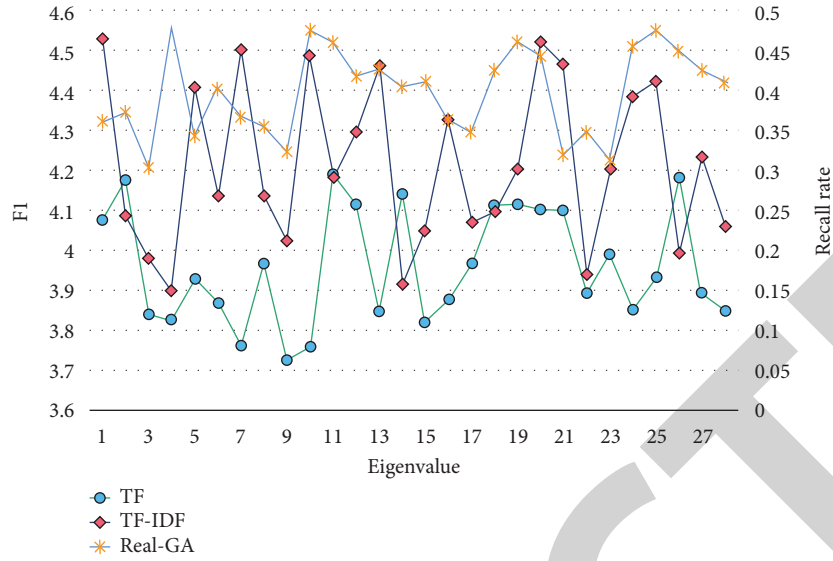

Figure 8: Comparison of F1 values of feature extraction algorithms.

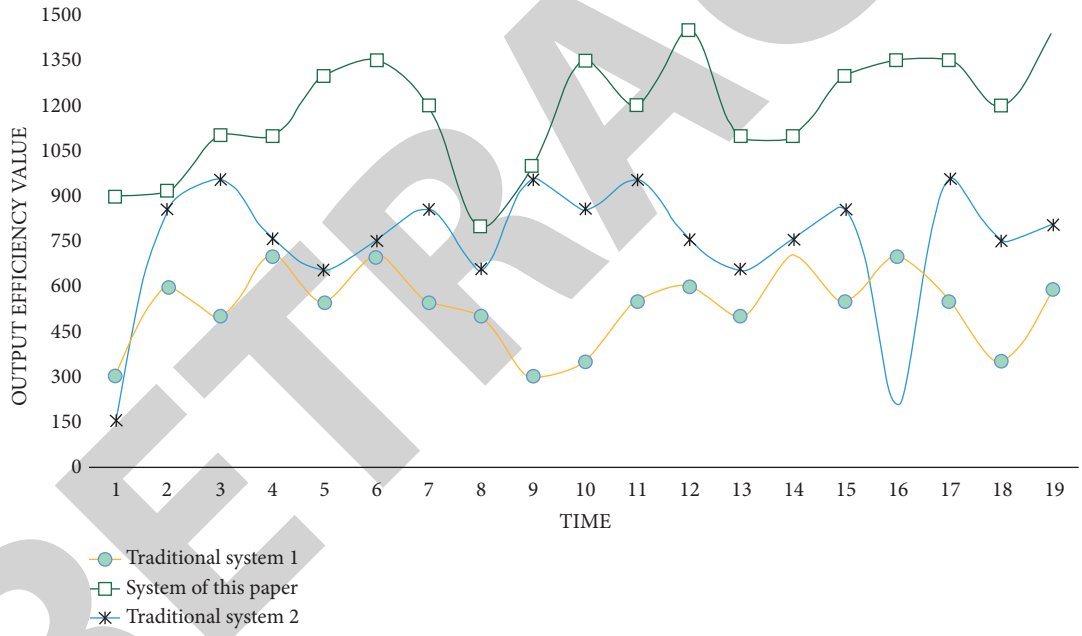

Figure 9: Comparison results of output efficiency values under different models.

the algorithm based on the machine learning model, we observe the change of the macro average curve with the increase of the feature quantity under different feature selection methods, as shown in Figure 7.

In order to further verify the accuracy of the model, the F1 values of the extracted features are compared, and the results are shown in Figure 8.

Obviously, the real encoding GA method proposed in this paper is more effective. In order to further test the operation detection efficiency of the model, it is compared with the system under other models, and the output efficiency value is obtained as shown in Figure 9.

As can be seen from Figure 9, the efficiency of this algorithm is the highest, while that of the traditional algorithm is poor. The experiment shows that the translation accuracy of this method reaches 94.2%, which is higher than that of the benchmark method by 39.5%. This shows that this method can reduce the influence of other factors, ensure the accuracy of cross-context English translation to a certain extent, and meet the requirements of improving the performance of English translation system. By analyzing the comparison curve of the training loss value of the algorithm data set, the stability of the model is good, and reliable identification data can be obtained.

5. Conclusions

This paper adopts the machine learning method to build a tree-like lexical semantic database for Chinese-English translation. According to the semantics in the tree lexical semantic database, the target is modified for structural automation. Support vector regression and the multilayer perceptron model are used to verify the quality of cross-context accurate English translation based on the machine learning model. The proposed k-means algorithm can not only obtain the correct number of clusters but also maximize the clustering effect. Experiments show that the translation accuracy of this method is 94.2%, which is 39.5% higher than the benchmark method. To a certain extent, it ensures the accuracy of cross-context English translation and meets the requirements of improving the performance of English translation systems. Automatic calibration of Chinese-English translation and subject word registration are realized, and the best semantic relevance feature quantity of each clause is calculated. The machine learning algorithm is used for automatic optimization to achieve automatic calibration of Chinese-English translation. The simulation results show that the accuracy of automatic calibration of Chinese-English translation using this method is high, and the relevance of translation calibration is strong.

Data Availability

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

Conflicts of Interest

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Yörük, A. Hürriyetoğlu, F. Duruşan, and C. Yoltar, "Random sampling in corpus design: cross-context generalizability in automated multicountry protest event collection," *American Behavioral Scientist*, vol. 66, no. 5, pp. 578–602, 2022.

[2] W. Van Atteveldt, M. A. C. G. Van der Velden, and M. Boukes, "The validity of sentiment analysis: comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms," *Communication Methods and Measures*, vol. 15, no. 2, pp. 121–140, 2021.

[3] J. Lv, J. Xiao, Q. Jia et al., "Identification of key pathways and genes in the progression of silicosis based on WGCNA," *Inhalation Toxicology*, vol. 13, no. 3, pp. 1–15, 2022.

[4] J. Zhang, Y. Tian, J. Mao, M. Han, and T. Matsumoto, "WCC-JC: a web-crawled corpus for Japanese-Chinese neural machine translation," *Applied Sciences*, vol. 12, no. 12, p. 6002, 2022.

[5] W. Gong, H. Chen, Z. Zhang et al., "A novel deep learning method for intelligent fault diagnosis of rotating machinery based on improved CNN-SVM and multichannel data fusion," *Sensors*, vol. 19, no. 7, p. 1693, 2019.

[6] B. Zhang, D. Xiong, and J. Su, "Neural machine translation with deep attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 154–163, 2020.

[7] T. Zhang, "A deep learning classification model for English translation styles introducing attention mechanism," *Mathematical Problems in Engineering*, vol. 2022, no. 5, Article ID 6798505, 77 pages, 2022.

[8] F. N. Al-Aswadi, H. Y. Chan, and K. H. Gan, "Automatic ontology construction from text: a review from shallow to deep learning trend," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 3901–3928, 2020.

[9] J. Su, Bo Zhang, and Da Xiong, "Alignment consistent recurrent neural networks for bilingual phrase embeddings," *Knowledge-Based Systems*, vol. 156, no. 8, pp. 95–119, 2018.

[10] D. Banick, A. Ekbal, and P. Bhattacharya, "Statistical machine translation decoding optimization pruning method based on machine learning," *IEEE Access*, vol. 2018, no. 6, pp. 77–95, 2018.

[11] N. Chatterjee and S. Gupta, "Efficient Phrase Table pruning for Hindi to English machine translation through syntactic and marker-based filtering and hybrid similarity measurement," *Natural Language Engineering*, vol. 25, no. 1, pp. 171–210, 2019.

[12] E. M. Ponti, H. O'horan, Y. Berzak et al., "Modeling language variation and universals: a survey on typological linguistics for natural language processing," *Computational Linguistics*, vol. 45, no. 3, pp. 559–601, 2019.

[13] D. Litman, H. Strik, and G. S. Lim, "Speech technologies and the assessment of second language speaking: approaches, challenges, and opportunities," *Language Assessment Quarterly*, vol. 15, no. 3, pp. 294–309, 2018.

[14] B. Deborah, "Automatic machine translation error recognition," *Machine Translation*, vol. 29, no. 2, pp. 52–76, 2015.

[15] S. Pouyanfar, S. Sadiq, Y. Yan et al., "A survey on deep learning: algorithms, techniques, and applications," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–36, 2019.

[16] S. Harrat, K. Meftouh, and K. Smaili, "Machine translation for Arabic dialects (survey)," *Information Processing & Management*, vol. 56, no. 2, pp. 262–273, 2019.

[17] D. J. M. D. Beatriz and D. Helena, "Automatic machine translation error recognition," *Machine Translation*, vol. 29, no. 1, pp. 1–24, 2015.

[18] N. Quang-Phuoc, V. Anh-Dung, and S. Joon-Choul, "The impact of word sense disambiguation on neural machine translation: a case study of Korean," *IEEE Access*, vol. 6, no. 3, p. 1, 2018.

[19] A. Poncelas, G. Winig, and A. Way, "An improved feature decay algorithm for statistical machine translation," *Natural Language Engineering*, vol. 28, no. 3, pp. 1–21, 2020.

[20] M. Araújo, A. Pereira, and F. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis," *Information Sciences*, vol. 512, no. 12, pp. 1078–1102, 2020.

[21] R. Peng, T. Hao, and Y. Fang, "Syntax-aware neural machine translation directed by syntactic dependency degree," *Neural Computing & Applications*, vol. 33, no. 23, pp. 16609–16625, 2021.

[22] S. Wehnert, S. Dureja, L. Kutty, V. Sudhi, and E. W. De Luca, "Applying BERT embeddings to predict legal textual entailment," *The Review of Socionetwork Strategies*, vol. 16, no. 1, pp. 197–219, 2022.