A Semi-supervised Learning-Based Method for Correcting Translation Accuracy of Literature Works

Hongzhu Jiang(✉)

Jilin University, Changchun, Jilin, China
2174962901@qq.com

Abstract. This paper studies the accuracy correction method of literary translation based on semi supervised learning, so as to improve the accuracy of Literary translation and reduce the translation error rate. Based on the word vector of recurrent neural network, the data preprocessing and feature extraction of translation of Literary works are realized by constructing a word alignment and segmentation model. Based on $TF - IDF$ algorithm, the translation grammatical features of Literary works are extracted, and K-means clustering algorithm is used to detect the accuracy features. Based on semi supervised learning, mistranslation features are identified, and translation accuracy correction of Literary works is realized through semantic feature analysis. The results show that this method can detect the mistranslation features in the grammar feature sample set. The number of mistranslation features detected is almost the same as the actual number of mistranslation categories in the corresponding data set, and the comprehensive detection performance is high; We can distinguish the mistranslated grammar from the correct grammar through grammar mistranslation correction, and the overall correction accuracy is higher than 98% .

Keywords: Semi supervised learning · Literary · Literary translation · Translation accuracy · Translation correction

1 Introduction

In the field of natural language processing, grammar correction has gradually become one of the main research directions. In the process of the continuous development of machine learning methods [1], more and more scholars began to study the application of machine learning algorithms to achieve grammar correction, so as to avoid the problems of low efficiency and poor accuracy of previous grammar correction. There is a great gap between the two language systems involved in English-Chinese translation in terms of expression and grammar, and it is impossible to achieve English-Chinese translation completely through literal translation. Generally, English-Chinese translation mostly adopts free translation and literal translation [2]. Due to the differences in the structure of the two language systems and the cultural factors involved, free translation has certain constraints and is prone to mistranslation [3, 4]. In addition, improper word selection

Z. Zhan et al. (Eds.): EIMSS 2022, AHCS 7, pp. 937–950, 2023.
https://doi.org/10.2991/978-94-6463-024-4_98

and lack of professional knowledge are also prone to grammatical mistranslation, which reduces the accuracy of English-Chinese translation and brings inconvenience to practical application. In order to effectively solve the above problems, it is necessary to select appropriate methods to accurately and efficiently correct the translation accuracy of English language Literary works, so as to improve the accuracy of English-Chinese translation [5].

A phrase corpus with a tag size of about 740000 English and Chinese words is constructed by improving the generalized maximum likelihood ratio detection, so that the phrase has a searchable function. By constructing the phrase structure through the phrase center, the part of speech recognition results can be obtained. According to the syntactic function of the parsing linear table, the English and Chinese structural ambiguity in the part of speech recognition results can be corrected, and finally the recognition content can be obtained. However, the effect is not ideal for grammatical problems such as lack of words and word order [6]; An online translation platform that can translate between multiple languages and English, complete the post translation sentence sorting, maintain the English knowledge base and other translation elements. It includes the architecture of five functional modules: memory maintenance, task management, manual correction, automatic translation and system receiving and sending. At the same time, it is divided into the system function realization process, which is mainly composed of integrating multiple languages and completing work on the basis of the project. However, the error correction process is time-consuming and the overall efficiency is not ideal [7].

Based on the above analysis, this paper studies a semi-supervised learning-based accuracy correction method for translation of Literary works. Based on the translation data preprocessing and feature extraction of English language Literary works, this paper uses semi supervised learning to identify the mistranslation features of English language Literary works, and constructs a mistranslation correction model by detecting the mistranslation features to correct the translation accuracy of English language Literary works. The experimental results show that this method can effectively and accurately correct all kinds of mistranslation problems in the translation of Literary works, improve the accuracy of translation, and provide convenience for users.

2 Design of Correction Method for Translation Accuracy of Literary Works

2.1 Translation Data Preprocessing and Feature Extraction

2.1.1 Word Vector Generation

Build a recursive neural network model to generate word vectors [8, 9], and digitally process English sentences to facilitate the understanding of natural language. By solving the non-fixed input vector x, the optional output vector y of the hidden layer h of the model is obtained, and the hidden layer h_t after the period t can be updated according to the following iterative formula:

$$h_t = h_{t-1}(\sigma + x_t) \tag{1}$$

Table 1. Word vector generation table

x_i	Initial h_i	w_i	e_i^x	Update h_i
What	h_1	1	h_1	$\vec{0}$
Are	h_2	0	$\vec{0}$	h_2
You	h_3	1	h_3	$\vec{0}$
Doing	h_4	1	h_4	$\vec{0}$
?	h_5	1	h_5	$\vec{0}$

Assuming that the connection weight matrices of the recursive model are V_w, W_h, U_h, W_w and U_w respectively, the outputs of the initial hidden layer at different times are h_i and h_{i-1} respectively, and the content of the source statement is represented by x_i, the following formulas are used to generate word vectors:

$$\begin{gathered} h_i = U_h h_{i-1} + tanh \times W_h x_i \\ \{\underline{w_i} = \sigma / U_w h_{i-1} - W_w x_i + V_w w_{i-1} \\ e_i^x = w_i / h_i \end{gathered} \tag{2}$$

Where, w_i is the control switch of the hidden layer output at time i, that is, the dynamic segmentation word. w_{i-1} and w_i is the representation of the previous moment and one digit of the control switch. The judgment basis of control switch is:

$$w_i = \begin{cases} 0, w_{i,1} \geq w_{i,2} \\ 1, w_{i,1} < w_{i,2} \end{cases} \tag{3}$$

With "What are you doing?" As an example, the generation process of word vectors is briefly described in Table 1. Make the last character 1 to ensure that the end of the statement is effectively output. Similarly, the target statement word vector e_i^y at time i is obtained.

2.1.2 Word Alignment Segmentation

In general, English language Literary works contain more long and difficult sentences, and longer English sentences will weaken the performance of the correction model. Therefore, a word alignment model is constructed to divide long sentences into multiple short sentences. According to the source statement x_i and the target statement y_i, the following optimization expression is used to obtain the best target statement:

$$y_i' = \arg max\{S(y_i \mid x_i)\} \tag{4}$$

It can be seen from the above formula that the calculation process of the maximum function can describe the search problem of generating sentences in the target sentence, and all possible sentences need to be obtained. Therefore, $S'(x_i \mid y_i)$ is defined as a statistical alignment model, and the expression is as follows:

$$S^{'}(x_i \mid y_i) = \sum S^{'}(x_i, \alpha_i \mid y_i) \tag{5}$$

Where α_i refers to any word in the source statement.

Make the probability distribution of all alignments consistent, and the alignment probability is not disturbed by word order. Set the start position and end position of statement segmentation as (l_1, j_1) and (l_2, j_2) respectively, where l represents the length of the source statement, j represents the length of the target statement, and the value ranges are $[l_1, l_2 - 1]$ and $[j_1, j_2 - 1]$ respectively. If the probability weight of the known statement pair is γ, and its adjustment parameter is β, the regularized alignment based on the statement length can be realized through the following expression:

$$\gamma = \beta \times \frac{1}{j_2 - j_1 + 1} + (1 - \beta) \tag{6}$$

Using the forward and reverse word alignment model, the target sentence probability $P(x_{l_1}^{l_2} \mid y_{j_1}^{j_2})$ and the source sentence probability $P(y_{j_1}^{j_2} \mid x_{l_1}^{l_2})$ are obtained, and the approximate joint probability is calculated. The probability matrix is constructed by coding language pairs. The matrix is divided into four quadrants by each segmentation point (l, j_1). For the start position (l_1, j_1) and end position (l_2, j_2) of sentence segmentation, two modes of monotonic alignment of the first and third quadrants and non monotonic alignment of the other two quadrants are designed. The calculation formulas of each alignment probability are as follows:

$$P_{l,j,1}(x_{l_1}^{l_2}, y_{j_1}^{j_2}) = P(x_{l_1}^{l}, y_{j_1}^{j}) \times P(x_{l+1}^{l_2}, y_{j+1}^{j_2}) \tag{7}$$

$$P_{l,j,0}(x_{l_1}^{l_2}, y_{j_1}^{j_2}) = P(x_{l_1}^{l}, y_{j+1}^{j_2}) \times P(x_{l+1}^{l_2}, y_{j}^{j_2}) \tag{8}$$

Traverse the segmentation points contained in the statement, and get the best segmentation point from the following formula:

$$(\hat{l}, \hat{j}, \hat{\delta}) = maxP_{l,j,\delta}(x_{l_1}^{l_2} \mid y_{j_1}^{j_2}) \tag{9}$$

Where, δ is an arbitrary constant with a value range from 0 to 1. The minimal partition statement length (l_{min}, j_{min}) is introduced to avoid the clause length being too short, and the length of the source statement and the target statement are limited to $[l_1 + l_{min} - 1, l_2 - l_{min}]$ and $[j_1 + j_{min} - 1, j_2 - j_{min}]$ sections respectively.

After the iteration cycle, the longest sentence length of the statement is no longer greater than the set statement length.

2.2 Extraction of Translation Features of English Language Literary Works

$TF - IDF$ (feature frequency inverse document frequency) feature extraction algorithm [10], whose essence is to calculate the occurrence frequency weight of a word and the similarity between texts, so as to obtain the optimal grammar of this word and achieve the purpose of extracting grammatical features. The algorithm can prevent the loss of semantics and vocabulary in grammar, and extract grammatical features with high accuracy. It can lay a solid foundation for feature extraction and correction of grammatical mistranslation.

Therefore, $TF - IDF$ algorithm is selected to extract grammatical features from the preprocessed translation data of Literary works to form a sample set of translation features of Literary works. The grammatical features are extracted by calculating the proximity of the text in the translation of Literary works and the weight of the frequency of words. The weight of words in the translation of Literary works can be obtained by the product of IDF and TF. On this basis, the best grammar of words is extracted. The formula is:

$$W(s_l, b) = IDF \times TF \tag{10}$$

Where, b represents the document number; s_l stands for vocabulary; W stands for weight. The task of IDF is to improve the criticality of words that occur less frequently and the difference between texts. Its operation formula is:

$$IDF = \lg \frac{M}{m_i} \tag{11}$$

Where, m_i represents the number of i words in all documents; M refers to the number of documents contained in the English-Chinese translation text. TF stands for characteristic frequency, and its expression is:

$$TF = \frac{\lg(S_F(s_l, d) + 1)}{\lg S} \tag{12}$$

Where, S represents the total number of words in document b after document processing; $S_F(s_k, d)$ represents the number of words s_l in document b.

2.3 Detection of Translation Accuracy Characteristics of Literary Works

As one of the common methods in the field of text detection [11, 12], the K-means clustering method essentially uses the centers of different classes of samples as the representatives of various classes to implement iteration, so as to continuously and dynamically adjust the centers of different classes to achieve clustering. Its advantages are strong adaptability and high autonomy. Its detection results can be updated automatically with the change of sample distribution mode, and the overall detection performance is high. Therefore, this paper chooses this method to detect the translation accuracy of Literary works.

Suppose that the extracted translation feature sample set of Literary works is $Y = \{y_i\}_{i=1}^{n}$, the specific clustering quantity parameter is expressed in k, and the sample set Y is clustered by the basic K-means; Evaluation function can be used to evaluate the inter class separation and intra class clustering of the clustering results. The evaluation index U can be selected as mean square deviation, and its operation formula is:

$$U = \sqrt{\frac{1}{n} \sum\nolimits_{j=1}^{k} \sum\nolimits_{q=1}^{n_j} \sum\nolimits_{p=1}^{d} (y_{j_q}^{p} - a_j^{p})^2} \tag{13}$$

Where, n and d respectively represent the sample quantity and dimension; a_j^p and $y_{j_q}^p$ represent the cluster center of class j and the p-th component of the sample; n_j represents the number of samples in class j in the clustering results. On this basis, by adjusting the value of the clustering parameter k, paying attention to the change of the indicator U, the appropriate clustering results are obtained, and based on this, the mistranslation features in the grammar feature sample set are detected. The detailed steps of translation accuracy feature detection of Literary works based on K-means clustering are as follows:

Step 1: let k and t be the initial clustering parameters and iteration parameters respectively, and $Y = \{y_i\}_{i=1}^{n}$ be the initial sample set of translation features of the extracted Literary works.

Step 2: after K-means clustering the initial sample set, obtain the initial clustering result $Y \to \{Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_{k^{(t)}}^{(t)}\}$. The clustering steps are as follows: ① randomly select $k^{(t)}$ samples from the initial sample set, and use the extracted samples to create the initial cluster center set, represented by $A^{(t)} = \{a_p{}^{(t)}\}_{p=1}^{k^{(t)}}$, where $a_p{}^{(t)}$ represents the cluster center. ② Calculate the similarity between all samples in the initial sample set and each cluster center, and divide each sample into the class where the cluster center with the highest similarity of the same sample is located. The similarity calculation formula is:

$$S(y_i, a_p^{(t)}) = \sum\nolimits_{p=1}^{d} (y_i^p - a_p^{(t)p}) \tag{14}$$

Where, the p-th component of sample y_i in the initial sample set is represented by y_i^p; The p-th component of cluster center $a_p^{(t)}$ is represented by $a_p^{(t)p}$. ③ Continue to update the cluster center $a_p^{(t)}$ of all classes after the update. The update expression is:

$$a_p^{(t)} = \sum\nolimits_{q=1}^{n_j} Y_{j_q}/n_j \tag{15}$$

The new cluster center set is obtained by updating formula (15). ④ Judge the cluster center set before and after the update. If they are different, go back to step ② and repeat the above process until they are the same; On the contrary, if the two are the same, the clustering results of this time can be directly obtained and expressed as $Y \to \{Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_{k^{(t)}}^{(t)}\}$.

Step 3: calculate the evaluation index $U^{(t)}$ based on the obtained clustering results, and compare the evaluation index $U^{(t-1)}$ and $U^{(t)}$ obtained from the previous round of clustering results. If the comparison result is $|U^{(t)} - U^{(t-1)}| < \varepsilon$, then:

$$\begin{cases} t = t + 1 \\ k^{(t+1)} = k^{(t)} + 1 \end{cases} \tag{16}$$

At the same time, return to step 2; otherwise, go directly to the next step.

Step 4: set the finally obtained clustering result as $Y \to \{Y_1^{(l)}, \ldots, Y_p^{(l)}, \ldots, Y_{k^{(l)}}^{(l)}\}$, and the number of samples in class $Y_p^{(l)}$ in this result is represented by $n_p^{(l)}$. Based on the value of $n_p^{(l)}$, arrange each class in the clustering result from low to high, and select the class with the lower number of samples at the top of the arrangement. The samples in these classes are the finally detected mistranslation feature samples. Complete the translation accuracy feature test of Literary works, and output the mistranslation feature samples.

2.4 Correct the Translation Accuracy of English Language Literary Works

2.4.1 Identifying Mistranslation Features of Literary Based on Semi Supervised Learning

Extract the translation data features that generate mistranslation of Literary works and input the semi supervised learning model [13, 14] to identify the mistranslation features of Literary works. Three sparse self encoders are used to form a semi supervised learning model, including input layer, hidden layer and output layer. The number of neurons in the model is constrained by sparse regularization terms, and the feature vector of translation data is extracted. Let the data set generating mistranslation of English language Literary works be $\{O_1, O_2, L, O_M\}$ and the normalized value of the data set be $h(O)$. Train each sparse self encoder separately and map the normalized value of the input translation data. The formula is:

$$N(O) = P(Q \times h(O)) + R \tag{17}$$

Where, $N(O)$ is the mapping value of the data set, P is the weight matrix of the encoder, and R and Q are the activation function and offset vector respectively. Minimize the cost function as the optimization objective of semi supervised learning model training, continuously adjust the value of model parameters until the optimal parameter value is obtained, and take $N(O)$ as the new feature expression of $h(O)$. Take $N(O)$ as the input of the next encoder, repeat the data mapping process, and take the output of the last encoder as the final feature expression of the translation data. Take all the new eigenvalues $\{r_1, r_2, L, r_U\}$ of the output layer of the semi supervised learning model as the new eigenvalue vector of the translation data, and all the weight values $\{s_1, s_2, L, s_V\}$ of the output layer as the original eigenvalue vector of the translation data. The range of weight values is [0,1], and the sum of ownership weight values is 1.

Preliminarily identify the mistranslation types of Literary works, map the two feature vectors to another vector with the same dimension, and the calculation formula is:

$$\begin{cases} \bar{r} = \frac{1}{r_u} \sum_{u=1}^{U} r_u \\ \bar{s} = \frac{S(X_{1,e} - X_{2,e}) \sum_{v=1}^{V} s_v}{V} \end{cases} \tag{18}$$

Where, r_u is the u-th new eigenvalue, s_v is the V -th original eigenvalue weight, $X_{1,e}$ and $X_{2,e}$ are the expected and observed values of the translation type respectively, S is the loss function of $X_{1,e}$ and $X_{2,e}$ [15, 16], and $\bar{r}$ and $\bar{s}$ are the mapping values of the eigenvectors respectively. Train a classifier for each mistranslation category of English language Literary works, and regard the mistranslation of English language Literary works belonging to this category as a positive example, and the mistranslation of other English language Literary works as a negative example. Calculate the probability value

Y_e of the e-th translation data generated by each classifier, and the formula is:

$$Y_e = \frac{T_e(y_1 + y_2 + y_3)}{Z_e + T_e} \tag{19}$$

Where, T_e and Z_e are the new feature vector and original feature weight vector of translation data respectively, and the membership degree of the mistranslation classifier of Literary works, and y_1, y_2, y_3 is the probability of mistranslation of grammar, proper nouns and long sentences of Literary works respectively. Compare the probability values of translation data output by all classifiers, take the translation data with the largest probability value as a positive example of the classifier, and regard the mistranslation category of Literary works corresponding to the classifier as the mistranslation type of Literary works generated by the e-th translation data. So far, the identification of mistranslation features of Literary works has been completed, laying the foundation for the next step of translation accuracy correction.

2.4.2 Translation Accuracy Correction Based on Semantic Features

Logistics model is a classic model in semantic feature analysis. By using this model to analyze the semantic features of Literary translation, we can get the semantic feature details of the pre correction target, and then complete the correction by means of text vocabulary matching method. As a chaotic model, logistics model has the characteristics of high randomness and sensitivity of initial features, and has the advantage of high environmental adaptability in translation accuracy correction, Building Logistic chaotic model with one-dimensional mapping:

$$x_{n+1} = \lambda x_n(1 - x_n) \tag{20}$$

Formula (20) represents a cluster attractor that can be translated in English, integrates the concept set of translation of English language Literary works, adaptively matches the context of translation of English language Literary works, and obtains the distribution model of characteristic concept set of translated text:

$$\begin{cases} \dot{x} = a + by = x^2 \\ \dot{y} = x \end{cases} \tag{21}$$

The function of the corrected attractor is:

$$\begin{cases} \dot{x} = -\sigma x + \sigma y \\ \dot{y} = -xz + rx - y \\ \dot{z} = xy - bz \end{cases} \tag{22}$$

In the sentence clustering feature extraction of the translation of Literary works, it is necessary to combine the semantic distribution differences in the translation of Literary works, rely on the chaotic attractor to cluster the semantic features, and establish a clustering model [17].

The clustering model will search for the most similar sentences to the translation of Literary works through the context feature matching and adaptive semantic variable method. At the same time, it will carry out automatic lexical feature analysis, and draw up the translation semantic code sequence of Literary works to be corrected. The size of the semantic distribution concept set is N, the semantic distribution concept set is x, and then the sentences similar to the wrong translation will be expressed as the column feature vector $x(n) \in R^N$ of $N \times 1$. Using the association semantic grouping expression algorithm, the clustering model description of Literary translation is obtained as follows:

$$x = \sum\nolimits_{i=1}^{N} s_i \psi_i = \psi_s \tag{23}$$

With the above model design, the semantic feature analysis and error correction of Literary translation are carried out under the chaotic model.

2.4.3 Design of Mistranslation Correction Process

Based on the model, the translation accuracy correction of Literary works is realized. Since each sub model in the model set is generated based on different mistranslation feature subsets, it is necessary to implement projection when inputting translation text data of Literary works into the mistranslation correction model set; After all the correction sub models output the correction results, the final correction results are obtained by voting based on the simple majority principle. The specific correction process is as follows:

(1) Input the mistranslation feature subset $\{C_1^{'}, C_2^{'}, \ldots, C_N^{'}\}$ and correction model set $E_{all} = \{e1, e2, \ldots, eN \times M \times B\}$ corresponding to each correction sub model;

(2) Let T represent the iteration time, then the translated text data of Literary works obtained in the previous iteration is expressed as $VTT - 1 = v_1^{T-1}, v_2^{T-1}, \ldots, v_n^{T-1}$;

(3) Each correction sub model $e_i \in E_{all}$, based on the mistranslation feature subset $C_i^{'}$ corresponding to the correction sub model e_i, projects the $VTT - 1$ vector and inputs it into the correction sub model e_i for correction;

(4) All sub models in the correction model set E_{all} are used to correct the $VTT - 1$ vector in turn, and vote after all correction results are counted;

(5) Output the final correction mark of the whole. When the mark is equal to -1, it means that there is a mistranslation of this grammar; When the tag is equal to 1, the syntax is correct.

3 Application Result Analysis

Taking the corpus of an English-Chinese translation platform as an example, part of the corpus is randomly selected as the experimental object. The selected experimental corpus contains 10 types of grammatical mistranslation, including verb errors, abbreviation errors, rhetorical errors, voice errors, word order, lexical errors, missing words, subject predicate errors and multiple words. This method is used to correct the grammatical mistranslation in the experimental corpus, Test the practical application effect of this method.

3.1 Preparation of Experimental Data

The experimental corpus is randomly divided into five data sets (a, B, C, D, e). The basic information of each data set is as follows: the total number of samples in data set a is 800, and the number of grammatical mistranslation samples is 90. The categories of grammatical mistranslation include abbreviation error, multi word, voice error, verb error, missing word, subject predicate error and vocabulary error; The total number of samples in the B data set is 550, and the number of grammatical mistranslation samples is 70. The mistranslation categories include subject predicate errors, verb errors, word order, missing words and voice errors; The total number of samples in the C dataset is 7000, and the number of grammatical mistranslation samples is 480. The mistranslation categories include missing words, multiple words, word order, lexical errors, singular and plural nouns, subject predicate errors, verb errors and abbreviation errors; The total sample number of D data set is 3000, and the sample number of grammatical mistranslation is 200. The mistranslation categories include verb errors, abbreviation errors, rhetorical errors, voice errors, word order, vocabulary errors, missing words and multiple words; The total sample number of e data set is 2000, and the sample number of grammatical mistranslation is 460. The mistranslation categories it contains include subject predicate errors, verb errors, abbreviation errors, rhetorical errors, voice errors, missing words, multiple words, word order, lexical errors, and singular and plural noun errors.

3.2 Result Analysis

Through this method, the experimental data sets are preprocessed, and the grammatical features are extracted. Then the mistranslation features in the extracted grammatical features are detected. The mistranslation feature detection results of this method are presented to test the effectiveness of this method. During the detection process, the initial clustering parameter k values of the five grammar feature sample sets (A1, B1, C1, D1, E1) extracted from the five experimental data sets are set to be 23, 25, 100, 45, 180 in turn. The method in this paper is used to repeat the experiments on each feature sample set for three times, and the average value is taken as the detection result of mistranslation features, as shown in Table 2.

Table 2. Statistics of mistranslation feature detection results of this method

Characteristic sample set number	Number of final clusters/piece	U value	Number of mistranslation features/piece
A1	35	0.424	2
B1	24	0.561	1
C1	115	0.416	10
D1	52	0.483	4
E1	184	0.220	8

By analyzing Table 2, it can be concluded that this method can realize the detection of grammatical mistranslation features. The number of grammatical mistranslation features detected in each grammatical feature sample set is very close to the number of mistranslation categories in the corresponding data set. It can be seen that this method has a good effect on the detection of mistranslation features.

In order to further test the mistranslation feature detection ability of this method and improve the reliability of the experimental results, this method is used to re implement 10 detection experiments on 5 grammar feature sample sets, and the comprehensive detection accuracy (PR), detection rate (DR) and F1 value of this method are counted to verify the mistranslation feature detection performance of this method. The statistical results are shown in Fig. 1.

Test the clustering convergence time in the process of detecting mistranslation features in each grammar feature sample set. Take three experiments as examples, and the test results are shown in Fig. 2.

It can be seen from Fig. 1 and Fig. 2 that the detection rate, accuracy and F1 value of the mistranslation feature of the method in this paper are high. The amount of data in the data set has little impact on the detection performance of the mistranslation feature of the method in this paper, and has a relatively large impact on the clustering convergence time in the detection process. The clustering convergence time of the method in this paper is relatively similar for the same sample set in the three experiments. On the whole, The comprehensive performance of this method is ideal.

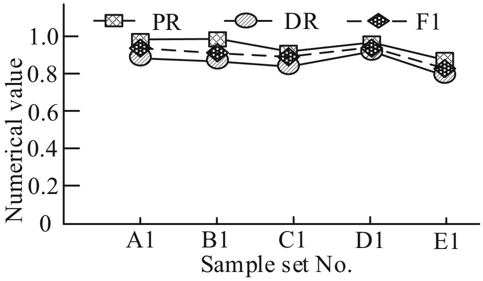

Fig. 1. Performance statistics of mistranslation feature detection of this method

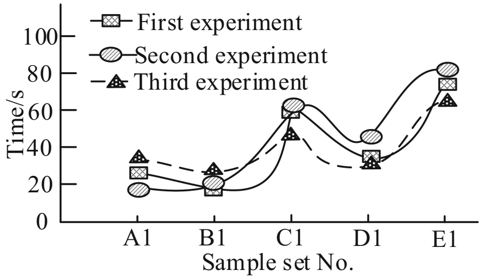

Fig. 2. Clustering convergence in mistranslation feature detection of this method

Table 3. Translation accuracy correction results of different algorithms

Correction algorithm	Translation accuracy correction results
Paper method	Total amount of English text resources integration
Improved generalized maximum likelihood ratio detection algorithm	Total amount of English text resources integration
English translation online platform	Total amounts of English text resources integration
Translation given by the teacher	Total amount of English text resources integration

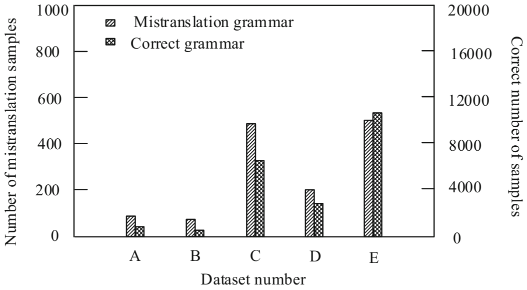

Fig. 3. Correction results of grammatical mistranslation of this method

In order to further prove the superiority of the proposed method, the improved generalized maximum likelihood ratio detection algorithm, the English translation online platform, and the correction results of the proposed method are compared with the translation given by the teachers. The translation of Literary works with errors: Total amounts of English text resources integration, as shown in Table 3.

It can be seen from Table 3 that the improved GMLR detection algorithm clears the complex s of the total amount, but does not clear the complex s of the resources. Similar to the online platform of English translation, the system does not clear the plural of the total amount, but the plural of resources. The proposed method corrects all errors in the translation of the Literary works correctly, and clears the plural of two words. The result is the same as the translation given by the teacher, and the recognition accuracy reaches more than 98%, which shows the superiority of the proposed method in the correction of translation accuracy.

Based on the above experimental results, this method corrects the syntax mistranslation in each data set, and the final correction results are shown in Fig. 3.

It can be seen from Fig. 3 that this method can correct the translation accuracy of Literary works. After correction, this method can effectively distinguish correct grammar from mistranslated grammar. The number of mistranslated grammar samples of the five experimental data sets corrected by this method is 89, 71, 476, 198 and 458, which is very close to the actual number of mistranslated grammar samples of each experimental data set, The correction accuracy of grammatical mistranslation can reach more than 98%, which shows that this method has high grammatical mistranslation correction performance, can correct the translation accuracy of Literary works with high accuracy, and the practical application effect is very ideal.

4 Conclusion

The quality of English-Chinese translation text has a direct impact on the application of scholars. Therefore, this paper proposes a semi-supervised learning-based correction method for the accuracy of translation of Literary works. Through preprocessing the collected translation data of Literary works, the TF-IDF algorithm is used to extract grammatical features from the preprocessed translation data of Literary works to form a sample set of grammatical features. The mistranslation features are identified through semi supervised learning method, and a mistranslation correction model is generated according to the detected mistranslation features to correct the grammatical mistranslation in the input translation text set of Literary works. The experimental results show that the proposed method can detect grammatical mistranslation features, and the number of grammatical mistranslation features detected in each grammatical feature sample set is almost consistent with the number of mistranslation categories in the corresponding data set. It has a high mistranslation feature detection rate, accuracy and F1 value, and has a good comprehensive detection effect; It can effectively distinguish mistranslated grammar from correct grammar through grammar mistranslation correction. The overall correction accuracy exceeds 98%, and the correction performance is stable and reliable. It has high practical applicability. It can reduce the mistranslation rate of English language Literary works, improve the accuracy and quality of translation, and provide help for scholars' efficient application.

References

1. Song G (2021) Accuracy analysis of Japanese machine translation based on machine learning and image feature retrieval. J Intell Fuzzy Syst 40(2):2109–2120
2. Gao J, Hua Y (2021) On the English translation strategy of science fiction from Humboldt's linguistic worldview—taking the English translation of three-body problem as an example. Theory Pract Lang Stud 11(2):186
3. Munawir M (2020) The kinds of translation error made by the students in writing abstract of the theses and dissertations. Sang Pencerah Jurnal Ilmiah Universitas Muhammadiyah Buton 6(2):58–66
4. Putri SR, Sujarwati I (2021) An analysis of syntactic translation error on communication students' abstract in Universitas Bengkulu. ETERNAL (English Teach J) 12(2):26–34
5. Schirmer A (2020) Aspects of the never-ending translation wars in South Korea: a cultural phenomenon and its reasons. Lebende Sprachen 49(5):390–410
6. Shasha D, Xiaotao G (2020) Design of intelligent recognition English translation model based on improved GLR algorithm. Comput Measure Control 28(4):161–164
7. Xinxin WANG (2021) Design of online platform for English translation in multi-language interactive environment. Microcomput Appl 37(10):70–73

8. Chen MY, Chiang HS, Sangaiah AK et al (2020) Recurrent neural network with attention mechanism for language model. Neural Comput Appl 32(1):7915–7923
9. Wu J, Hu C, Wang Y et al (2020) A hierarchical recurrent neural network for symbolic melody generation. IEEE Trans Cybern 50(6):2749–2757
10. Li J, Dai L (2020) Multiple key information extraction simulation based on integer linear simulation. Comput Simul 37(10):365–368,383
11. Muhima RR, Kurniawan M, Pambudi OT (2020) A LOF K-means clustering on hotspot data. Int J Artif Intell Robot (IJAIR) 2(1):29
12. Sinambela Y, Herman S, Takwim A et al (2020) A study of comparing conceptual and performance of K-means and fuzzy C means algorithms (clustering method of data mining) of consumer segmentation. Jurnal Riset Informatika 2(2):49–54
13. Pan J, Wong TM, Wang H (2022) Navigating learner data in translator and interpreter training: insights from the Chinese/English translation and interpreting learner corpus (CETILC). Babel 68(2):236–266
14. Zhu QY, Li TT (2020) Semi-supervised learning method based on predefined evenly-distributed class centroids. Appl Intell 50(9):2770–2778
15. Mei J, Gao B, Xu D et al (2020) Semantic segmentation of 3D LiDAR data in dynamic scene using semi-supervised learning. IEEE Trans Intell Transp Syst 21(6):2496–2509
16. Ke J, Gong C, Liu T et al (2020) Laplacian Welsch regularization for robust semisupervised learning. IEEE Trans Cybern 52(1):164–177
17. Yang M, Liu S, Chen K et al (2020) A hierarchical clustering approach to fuzzy semantic representation of rare words in neural machine translation. IEEE Trans Fuzzy Syst 28(5):992–1002