International Journal of Data Science

ISSN online: 2053-082X - ISSN print: 2053-0811
https://www.inderscience.com/ijds

Study on English machine translation based on feature extraction algorithm and big data information technology

Zheng Chao, Yixun Lin, Xingzu Zhan

DOI: 10.1504/IJDS.2025.10072978

Article History:

Received:	27 May 2025
Last revised:	17 July 2025
Accepted:	23 July 2025
Published online:	16 January 2026

Study on English machine translation based on feature extraction algorithm and big data information technology

Zheng Chao

School of Big Data and Basic Sciences,
Shandong Institute of Petroleum and Chemical Technology,
Dongying, 257061, Shandong, China
Email: zsw5483035@163.com

Yixun Lin*

School of Economics,
Jinan University,
Guangzhou, 510000, Guangdong, China

and

YX Tech Co., Ltd.,
Guangzhou, 518000, Guangdong, China
Email: 19874459335@163.com
*Corresponding author

Xingzu Zhan

YX Tech Co., Ltd.,
Guangzhou, 518000, Guangdong, China

and

College of Electronics and Information Engineering,
Shenzhen University,
Shenzhen, 518000, Guangdong, China
Email: 20222270054@email.szu.edu.cn

Abstract: The proposed intelligent automatic English translation system leverages advanced feature extraction algorithms and big data technologies to enhance translation accuracy and efficiency. Central to this system is an N-Gram-based scoring model, which evaluates translation quality by analysing word sequences. This model is further refined through the development of an English corpus scoring framework, enabling more precise assessments. Incorporating Latent Dirichlet Allocation (LDA), the system employs weighted LDA indices to assess the semantic depth of translations. When these indices are well-aligned, they indicate a translation that captures the nuances and depth of the original text. Conversely, scattered LDA indices suggest a loss of key semantic elements during translation. The integration of behavioural

decompression algorithms facilitates the optimisation of translation processes, ensuring that the system delivers high-quality English-Chinese translations by effectively capturing and preserving semantic information.

Keywords: feature extraction; big data information technology; English-Chinese translation; interactive.

Reference to this paper should be made as follows: Chao, Z., Lin, Y. and Zhan, X. (2025) 'Study on English machine translation based on feature extraction algorithm and big data information technology', *Int. J. Data Science*, Vol. 10, No. 7, pp.224–241.

Biographical notes: Zheng Chao got the Doctor's degree from Lyceum of the Philippines University in 2023. His research direction was translation studies and language studies. After graduating from Lyceum of the Philippines University, he has been working at the School of Big Data and Basic Sciences, Shandong Institute of Petroleum and Chemical Technology. His research interest mainly covers English teaching and translation studies. He taught college English and translation course as well. Based on the teaching experience, he has published one book and several high-quality papers related to translation and English teaching.

Yixun Lin is the CEO of YX Tech Co., Ltd. Currently, he is pursuing a degree in Economic Statistics at Jinan University. His research focuses on areas such as privacy computing and artificial intelligence. Additionally, he has received numerous national, provincial and municipal awards in the field of academic research.

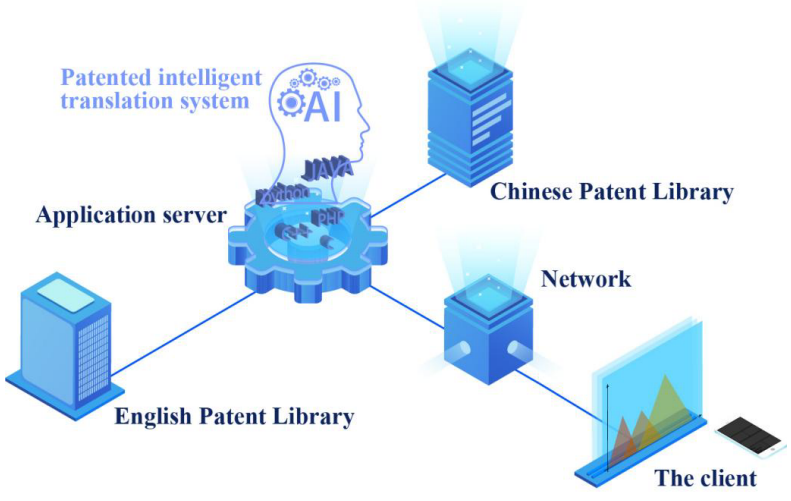
Xingzu Zhan is the CTO of YX Tech Co., Ltd. He entered the Academician Class of the School of Electronic and Information Engineering at Shenzhen University in 2022 to pursue a Bachelor's degree. He is currently focused on research in artificial intelligence, blockchain technology, and their interdisciplinary applications.

1 Introduction

Big data is a new information technology that has been applied extensively in a variety of industries, including government governance, healthcare, education, and agriculture. Big data technology, along with other science and technology, has made significant advancements in a variety of sectors, advanced technological advancements, and mechanism upgrades, and brought about profound changes in all spheres of existence. Big data information technology refers to the application technology of big data, covering various big data platforms, big data index systems, and other big data application technologies. Big data refers to a collection of data that cannot be captured, managed, and processed with conventional software tools within a certain time range. It is a massive, high-growth, and diversified information asset that requires a new processing mode to have stronger decision-making power, insight, and process optimisation ability. This paper discusses the application of big data in the English automatic translation system. English, as the main language of international communication, has received more and more attention, and in the process of transnational communication, the English translation

system is particularly important (Wang et al., 2020). With this as the background, this paper adopts the method of constructing Scoring rules of the English corpus and the N-gram scoring model. Figure 1 is an intelligent translation system (Yang and Yang, 2020; Zhou and Zhang, 2020).

Figure 1 Intelligent translation system (see online version for colours)



2 Literature review

In the information age, emerging information technology is flooding into all aspects of society at a rapidly changing speed. In recent years, emerging information technologies such as artificial intelligence, the Internet of Things, big data, and cloud computing have been widely used in various fields of society. Among them, the emergence of big data is changing people's conventional statistical method – probability calculation based on partial generalisation. Big data clarifies both known and unknown information. The wide application of big data information in various fields has fully proved its value. However, in various fields, there are great differences in the entry point, focus, and foothold of the application of big data technology. At present, the active application in the field of automatic translation scoring is mainly to score machine translation and to measure the quality of machine translation models based on different algorithms, but the application in automatic translation scoring is not very widespread (Bai, 2021; Wang et al., 2020). On the one hand, the automatic scoring model of machine translation cannot be applied to the automatic scoring of CET-4 and CET-6 because the length of CET-4 and CET-6 translation is less than 150 words (Li et al., 2021). So the automatic scoring model of composition cannot be directly applied to the automatic scoring of CET-4 and CET-6 translation (Chai, 2021). At present, the automatic translation scoring system is mainly based on the following three types. SER system uses this method. Based on edit distance, which is often used to measure how similar two strings are, the minimum number of operations (increase, delete, insert, and replace) required to change one string into the other (Li, 2020). The smaller the editing distance, the higher the quality of the

corresponding translation (Ajitha et al., 2020). Hou et al. proposed an automatic evaluation model for machine translation based on test points (Hou et al., 2020; Wang et al., 2021). Li and Geng (2020) brought semantic factors into the automatic scoring model of Chinese translation; Thotapalli et al. extracted text features from four directions of translation language semantic connection inertia and test points and finally established a translation scoring model optimised. Compared with the translation scoring model based on multiple linear regression, the model improved relevance by 6% (Thotapalli et al., 2021; Elouariachi et al., 2020). The Bingguo English composition intelligent rating system and the correcting website composition automatic rating system developed by Ouariachi et al. have been put into practical application and can give feedback information (Ouariachi et al., 2022; Jiang et al., 2022). Gan et al. (2021) used the N-Gram model to calculate the probability of sentence rationality.

Both of these two subjective scoring techniques are becoming more and more mature. Therefore, the automatic scoring method of human translation can be realised on this basis (Mintorini and Mahmud, 2020).

3 English automatic translation scoring model based on N-Gram

3.1 Basic principle of apriori algorithm

Big data technology has the advantage of being able to gather disparate data and create a database with the aid of contemporary network and communication technologies. Data connection policies are typically deleted using a priori methods. It is often used to find the dataset by the data value, and knowing the structure of these packages can help us determine, thus saving costs and increasing financial efficiency (Fei and Tian, 2020; Li et al., 2020). If there are four commodities: Commodity 0, commodity 1, commodity 3, and commodity 4, and you want to know which combination of commodities are commonly purchased (possibly at the same time), you can see the following combination form by exhaustive method.

To find collections of items (itemsets) that are often purchased together, you need to calculate support in the above description. For example, for the set {0,2}, when calculating its support, we need to traverse every record and query whether the record contains 0 and 2. If it contains 0 and 2, the value count is added by 1. The first step is to calculate the occurrence times of each member item respectively, called support degree. By scanning the database for the first time, the results are shown in Table 1.

Table 1 Degree of support for various projects

Project	*Support degree*
{1}	3
{2}	6
{3}	4
{4}	5

In the next step, the list of project pairs composed of all frequent items in the item set is shown in Table 2. For example, {1,2} indicates that 1 and 2 are in a transaction at the

same time (Kim et al., 2020; Boztas and Tuncer, 2021). It is not difficult to calculate all 2 item sets and their support degree:

Table 2 2 projects aggregate their support

Project	*Support degree*
{1,2}	3
{1,3}	1
{1,4}	2
{2,3}	3
{2,4}	4
{3,4}	3

Next, all three project sets and their associated support can be counted, as shown in Table 3.

Table 3 Projects aggregate their support

Project	*Sustain degree*
{1,2,3}	1
{1,2,4}	2
{1,3,4}	1
{2,3,4}	2

3.2 N-Gram text feature extraction

The resources integrated into big data information technology are endless, and the knowledge capacity brought by it is also extremely huge. Most machine learning algorithms require data processing, which means that text can be represented as a series of feature vector sets. When a word is present in the designated training text, its corresponding Boolean attribute value is set to 1; when it is not, it is set to 0. All of the feature vector's attribute values are set to 0 in the initial state, allowing each text to be represented as the collection of words that make up the text (Fang and Wang, 2024; Meng et al., 2023).

1 *Experiments settings*

To demonstrate the efficacy of the Apriori algorithm in automatically extracting text features, the corpus gathered for this paper is short, covering only three themes, and each text is no more than 150 words.

2 *Experimental result*

We measure the effect of text classification based on word set feature vector based on word bag feature vector and n-gram feature vector from three measurement indexes: recall rate accuracy and F1 measure. First, we investigated the experimental results using N-gram text features, as shown in Table 4.

Table 4 Reuters news text classification results based on N-gram features

TF and DF	*N*	*Recall rate*	*Precision*	*F1*	*Number of features*
TF:5	1	77.23	83.56	80.27	9674
DF:3	2	80.35	82.04	81.19	28,046
	3	77.57	82.75	80.08	38,647
	4	78.19	82.32	80.18	45,877
TF:5	1	77.18	83.66	80.28	6333
DF:10	2	80.06	82.07	81.05	13,599
	3	77.97	82.28	80.08	17,709
	4	78.22	82.14	80.13	20,469
TF:10	1	76.93	83.98	80.31	4069
DF:20	2	79.07	82.05	80.53	7067
	3	77.33	82.68	79.92	8758
	4	76.99	82.92	79.85	9908

To observe the experimental results more intuitively, we drew three line charts for different values of TF and DF. Where the ordinate is recall rate accuracy or F1 measure (Shorthand for R/P/F1); The abscissa is the value of N, as shown in Figures 2–4.

Figure 2 (T)F = 5, DF = 3 (see online version for colours)

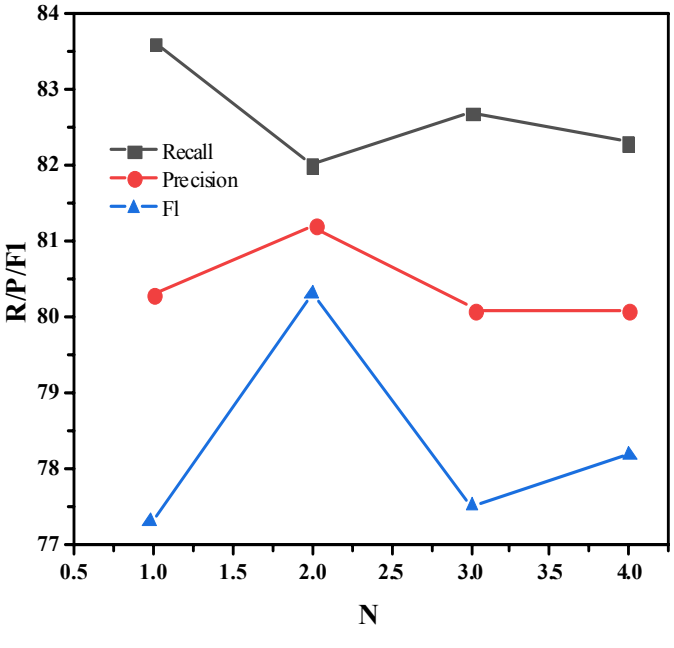


Figure 3 (T)F = 10, DF = 5 (see online version for colours)

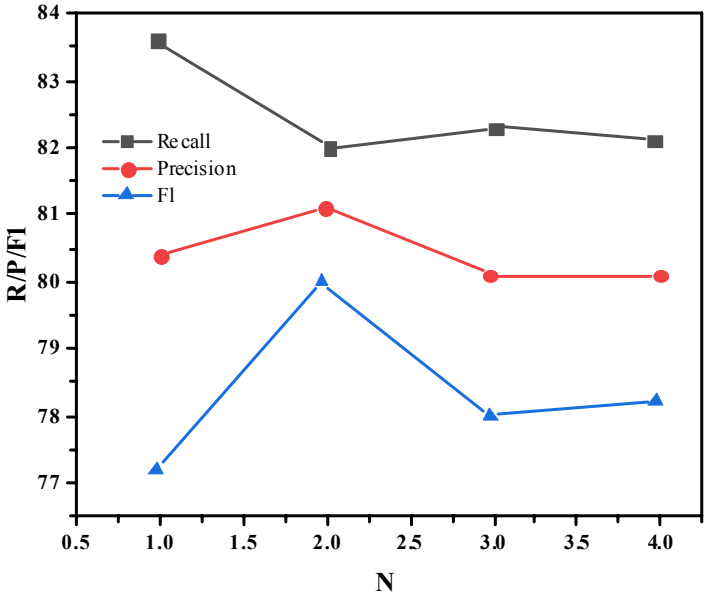
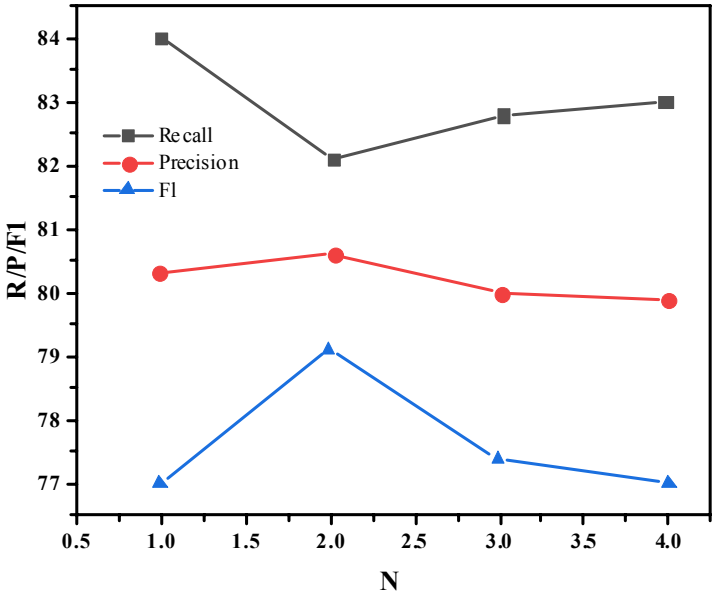


Figure 4 (T)F = 10, DF = 10 (see online version for colours)



The three figures above demonstrate that, for $N = 2$, the classification accuracy is highest when unary text features are used to represent text, but the classification recall rate and F1 measure are highest when binary text features are used to represent text. Overall, using binary text characteristics to represent text has many benefits (Pan et al., 2021; Yao et al., 2021).

4 Intelligent automatic English translation system based on feature extraction algorithm

Big data information technology can handle large amounts of data. Besides, it can process different types of data. Data information technology can not only process some large and simple data but also process some complex data, such as text data, sound data, image data, etc. As an auxiliary tool for English-Chinese translation, it primarily uses word segmentation as the basis for translation and connects the segments to form the final translated result. Although China has begun to optimise the framework and program development of the translation system, few achievements have been achieved, and a perfect translation system has not been formed (Wang et al., 2023; Tian, 2023).

4.1 Introduction of feature extraction algorithm

With the development of science and technology, the application of information technology has become more and more extensive. This paper introduces the feature extraction algorithm and big data information technology, extracts the mapping of the best context into the translation process through the feature extraction algorithm, completes the standard extraction of feature context, and describes the extracted best context through the semantic ontology mapping model. Note that there are N meanings at the time of translation, including the k-type translation, the number of the Ni meanings ($i = 1, 2, \ldots, K$), and the resulting test of k-type semantic definition $Xi = \{Xi1, Xi2, \ldots, Xi\mathrm{N}\}$, $X = i = 1,2, \ldots, K$; $j = 1,2, \ldots, Ni\}$ is the result of a directional n-dimensional vector. The basic definition process can be accomplished by the following limitations:

$$\alpha_i = \frac{1}{N}\sum_{j=1}^{N_i} x_{ij} \tag{1}$$

α_i is the best context α selection process is:

$$\alpha = \frac{1}{K}\sum_{i=1}^{K_i} \alpha_{ij} \tag{2}$$

$$S_w = \sum_{i=1}^{K}\sum_{j=1}^{K_i}\left(\alpha_{ij} - \alpha\right)\left(\alpha_{ij} - \alpha\right)^K \tag{3}$$

$$S_B = \sum_{i=1}^{K}\left(\alpha - \alpha_{ij}\right)\left(\alpha - \alpha_{ij}\right)^K \tag{4}$$

λ is set to the optimal point of the semantic point matrix association matrix, and since f is the standard for measuring the degree of semantic point correlation, the value of α can be directly related to process relationships. The mean definition of the organisation matrix has the highest of the $K - 1$ mean definitions, and since the definition resulted from this is R ($R \le K - 1$), the mean of the language of the features in the positive context can be shown as follows β:

$$\beta = [\alpha_1, \alpha_2, ..., \alpha_R] \tag{5}$$

4.2 Construction of semantic text model for automatic English translation

To use automatic English translation, an automatic English translation has been developed that combines translator and semantic analysis, and the semantic ontological model of non-English translation is manually created. Assuming that the distribution of English translations is A = O = {C, HC, R, I, A}, the vague diagram of the translation is:

$$\theta : S \to S \times [-0.5, 0.5] \tag{6}$$

$$\theta(s_i) = (s_i, 0), s_i \in S \tag{7}$$

The English translation is defined as:

$$O < C, I, H^C, R, A \tag{8}$$

and:

$$O' < C', I', H^{C'}, R', A' \tag{9}$$

In the above two formulas, the semantic feature extraction method is adopted to design concept lattice allocation, and the fuzzy reasoning method is adopted to obtain the parameter correlation parameter set of Automatic English translation:

$$\Delta : [0, T] \to S \times [-0.5, 0.5] \tag{10}$$

Namely:

$$\Delta\beta = s_k, K = round(\beta) \tag{11}$$

$$\Delta\beta = a_k \tag{12}$$

$$a_k = \beta - k, a_k \in [-0.5, 0.5] \tag{13}$$

Based on the correlation semantic mapping method, the binary semantic fusion feature parameter distribution of English translation is:

$$(\bar{s}, \bar{a}) = \omega_2 \left((s_1, a_1), \left(\omega_1, a_1'\right) \right), \left((s_2, a_2), \left(\omega_2, a_2'\right) \right), ..., \left((s_n, a_n), \left(\omega_n, a_n'\right) \right) \tag{14}$$

$$(\bar{s}, \bar{a}) \quad = \Delta \left(\frac{\sum_{j=1}^{n} \Delta^{-1} \left(\left(\omega_j, a_j'\right) \Delta^{-1} \left(s_j, a_j\right) \right)}{\sum_{j=1}^{n} \Delta^{-1} \left(\omega_j, a_j'\right)} \right) \tag{15}$$

$$\begin{aligned} &\Delta \left(\frac{\sum_{j=1}^{n} \Delta^{-1} \left(\left(\omega_j, a_j'\right) \Delta^{-1} \left(s_j, a_j\right) \right)}{\sum_{j=1}^{n} \Delta^{-1} \left(\omega_j, a_j'\right)} \right) \\ &= \Delta \left(\frac{\sum_{j=1}^{n} \beta_j \beta_j'}{\sum_{j}^{n} \beta_j'} \right) \end{aligned} \tag{16}$$

$$\left(\overline{s},\overline{a}\right)=\Delta\left(\frac{\sum_{j=1}^{n}\beta_j\beta_j^{'}}{\sum_{j}^{n}\beta_j^{'}}\right) \tag{17}$$

Among them:

$$\sum_{j=1}^{n}\omega_j=1,\overline{s}\in S,\overline{a}\in\left[-0.5,0.5\right] \tag{18}$$

Establish semantic evaluation index of English translation:

$$I_jI\left(j=1,2,...,n\right) \tag{19}$$

4.3 System software development and design

1 *Process*

The English automatic translation system has undergone more significant alterations as a result of the integration of big data information technology and feature extraction algorithms (Luckhoo and Peer, 2023). The following software development core process is designed, and the system's software development is based on processing between English and Chinese, with the mapping object extraction setting of the best translation context serving as the core (Dong, 2024; Meghanathan, 2024).

Step 1: system initialisation, focusing on short document collection information; *Step 2*: set up the section, mainly with the vector space of the subject word; *Step 3*: call system database to extract semantic information; *Step 4*: set up cyclic functions, including hypothesis propositions and translation criteria; *Step 5*: Judge whether the current translation meets the hypothesis. If it meets the translation criteria, enter metric judgement; otherwise, re-translate the translation; *Step 6*: Use the measurement method to judge whether the semantics in the current translation results are consistent with the original text, and give the similarity measurement results; *Step 7*: Based on the measurement results, extract the statement with the largest measurement from it and take it as a mapping object to take mapping processing to the translation statement and generate the mapping translation; *Step 8*: Count segmentation words in translation as test objects of mapping translation; *Step 9*: Judge whether the current translation is consistent with the context of the original text. If so, output the translation result; otherwise, return to the third step to extract semantic information again (Zhang, 2024).

2 *Program design*

Modify the system software running program, apply initialisation treatment to brief documents, and produce particular items by the system software development process. Second, call the system database by the translation requirements after extracting semantic information. Modify the system software's core program by this design concept:

```
BEGIN

WordFormalt = NEWemanticfuzzy;//System software initialization processing,
with short document collection as the processing object
greycorrelation = NEWsemantictopic//Set up the entries, including the subject
word vector space
```

Word = Englishtranslation (WordList)//Extracting semantic information from the system databaseWHILE (semanticcontextISNOTNULL){//Set loop function IF (Wordemanticfuzzyoptimal){//Hypothesized propositions are set as cyclic control conditions

Word = fraturematching (information);//Set up the processing scheme of the opposite condition of the hypothesis proposition, that is, the current term is not in the thesaurus, then the whole search is adopted, Adjust the coefficient to complete the turn ELSE

Information (reasonable);//Judge the rationality of the current translation results in Sim = matchingcalculation (Word, Node);//Using the measurement method, the consistency of the semantics in the current translation results is judged and the similarity measurement results are given.

SimList.put (lauygdbgfOf (Sfgr));//Extract the statement with the largest metric and take it as a mapping object Word. rfregvface (If (simregrfist).Nbtjuke);//Taking word segmentation in translated text as a statistical object, statistical results are generated.

END

By running the above program, short documents can be translated, and the translation items and semantics can be set according to the translation requirements. For inconsistent sentences or sentences with poor consistency, the sentences with the maximum consistency are extracted from the translation as a reference, and the semantic standards of the translation are defined through mapping. According to this standard, word segmentation is counted. Then run this program again to take a second translation process, to achieve more accurate translation results.

4.4 System test analysis

1 *Test parameter*

To verify that the system software development scheme proposed in this paper can meet the requirements of interactive English-Chinese translation, the system functions are tested in this study. The test parameters are:

1 Phrase translation volume: 300 char.

2 Semantic recognition efficiency: 20kbit/s.

3 Translation rate: 13kbit/s.

4 Short passage translation: 450 words.

In this experiment, a random extraction method is adopted to set the translation text, and the text content within the set parameter range of phrase and short passage translation volume is randomly extracted from the database.

1 *Test Number 1*: Two types of context translation are set, and the comprehensive revision parameter is limited to 0.1×10^{-3}.

2 *Test Number 2*: Three kinds of context translation are set, and the comprehensive revision parameter is limited to 0.2×10^{-3}.

3 *Test Number 3*: Two types of context translation are set, and the comprehensive revision parameter is limited to 0.1×10^{-3}.

4 *Test Number 4*: There are at most 3 kinds of contextual translation, and the comprehensive revision parameter is limited to 0.2×10^{-3}.

5 *Test Number 5*: Two types of context translation are set, and the comprehensive revision parameter is limited to 0.1×10^{-3}.

6 *Test Number 6*: Three kinds of context translation are set, and the comprehensive revision parameter is limited to 0.2×10^{-3}.

3 *Test method*

First: correlation test

The semantic correlation degree of the translation system largely determines the quality of English-Chinese translation, so the correlation degree is selected as the test index. The proposed translation system was used as the experimental group to translate texts randomly selected from 6 experiments. The correlation score ranges from 1 to 5, and the higher the score, the higher the correlation between the translated text and the original text.

Second: weighted Latent Dirichlet Allocation (LDA) index test

This test uses the weighted LDA index as the test index and obtains the correlation of translation semantics between the traditional translation system and the translation system in this study through a simulation test. If the LDA index can be connected in an orderly manner, it is considered that the translation results generated by the translation system have a strong semantic correlation and grasp the semantic focus of the text. If the LDA index distribution is scattered, the translation results generated by the translation system are considered to have weak semantic correlation and low translation accuracy. According to the setting of test times, the weighted LDA index in the 6 groups of tests was tested respectively.

4 *Test result*

According to the test method, the Chinese texts of the six groups of tests were translated, and the correlation degree of the translation was scored by experts. The results are shown in Table 5. The statistical results in Table 5 show that the performance of the translation system designed in this study is higher than that of the traditional translation system, and the translated version has a higher correlation. Except for the third test, which got 4 points, the correlation degree of the other five tests reached the highest score.

As per the experimental methodology, the text content of the experiment was translated six times using both the conventional translation system and the translation system created for this study. The corresponding weighted LDA index values were then produced. The distribution of the weighted LDA index in traditional translation systems

is relatively scattered, which shows that the correlation between vocabulary and semantics in translation results generated by this translation system is not very strong. It can be judged that the system scheme has improved the relevance of lexical meaning (Kim et al., 2020).

Table 5 Correlation test results

Test number	*Types of contextual translation*	*Traditional translation system correlation rating*	*The correlation degree of the translation system in this study was scored*
1	2	3	5
2	3	3	5
3	2	4	4
4	3	2	5
5	2	3	5
6	3	3	5

4.5 Interactive English-Chinese translation simulation experiment

1 *Parameter determination*

Experimental parameters are set, in Table 6.

Table 6 Test parameter

Basic parameter	*Data value*
Phrase translation	300
Short translation	450
Translation	12
Semantic recognition rate	20

The experiment designed in this paper requires a random selection of experimental objects. Certain conditions and restrictions on experimental objects are shown in Table 7.

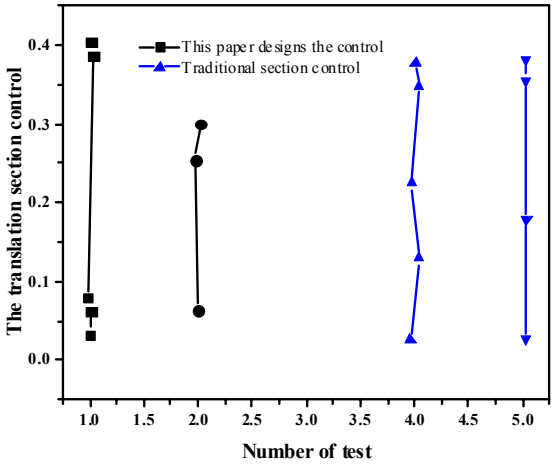
Table 7 Set up test data

Test number	*Types of contextual translation*	*Comprehensive revision parameter*
1	2	0.1×10^{-3}
2	3	0.2×10^{-3}
3	2	0.1×10^{-3}
4	3	0.2×10^{-3}
5	2	0.1×10^{-3}
6	3	0.2×10^{-3}

2 *Outcome analysis*

Figure 5 displays the number of node control points in the translation process. The translation control distribution map created in this paper is shown on the left, indicating that the distribution is comparatively balanced. The traditional English-Chinese translation system's translation control distribution is shown on the right. Restrained distribution might reflect the link between semantics and context of the translation system (Boztas and Tuncer, 2021; Fang and Wang, 2024). The feature extraction algorithm-based interactive English-Chinese translation system designed in this paper has a compact distribution of translation node control points without loose distribution, indicating a high translation accuracy.

Figure 5 Comparative test results (see online version for colours)



The comparison of weighted LDA indices is shown in Figures 6 and 7.

Figure 6 Weighted LDA index comparison Diagram 1

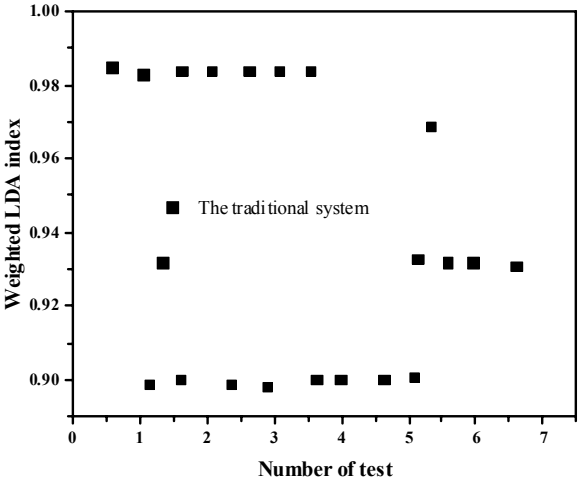
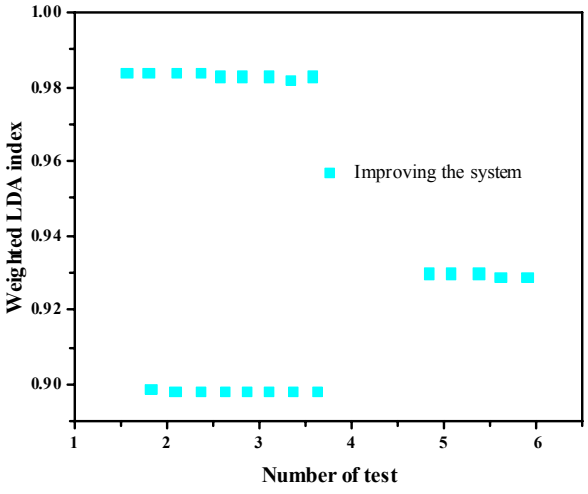


Figure 7 Weighted LDA index comparison Diagram 2 (see online version for colours)



The analysis of the figure above shows that the weighted LDA indices of the English-Chinese interactive translation system based on the feature decompression algorithm developed in this paper can be coordinated and distributed in an orderly manner, resulting in a traditional English translation. There is no correlation between the weighted LDA index in the Chinese translation system (Meng et al., 2023). The unweighted LDA index is a measure of semantic depth connection in the translation process. When weighted LDA indexes are connected in an orderly manner, it indicates that the translation process is vivid and deep; when weighted LDA indexes are scattered, it indicates that the key points of translation semantics are not grasped.

5 Conclusion

Big data information technology has brought value to data scattered around the corners of social production and life. Big data technology summarises, analyses, and changes human society in the form of quantitative data, thus making human society develop towards digitalisation, automation, and intelligence. When a big data stream collides with other technologies, it will burst out with greater power. Therefore, the application of big data in various fields marks only a beginning, with no foreseeable end, and holds limitless potential for the future. The automatic English translation system makes a detailed analysis of English lexical features using semantic analysis, and words using the semantic fuzzy matching phrase automatic analysis method. This paper describes the English-Chinese translation based on a special decision algorithm. To select the semantics of the features, a character extraction algorithm is introduced, a standard semantic ontology mapping is developed, and a good solution for the English-Chinese translation algorithm is selected. Finally, the English-Chinese translation process is coded. In conclusion, the development of an automatic English translator based on a combination of translator intelligence and phrase translation can improve the intelligence and automation of translation. The design process has two main components: software system design and translator algorithm design. The combination of content analysis and translation has

enhanced the automatic translation system, and the software created by the automatic translation is shown appropriately. The test's outcomes demonstrate the system's high degree of automation, speech, interpretation, and intelligence.

Funding

Supported and funded by Key Project of Art Studies in Shandong Province, titled A Study on the Translation and International Communication of Lyu Opera from the Perspective of Communication Studies (No.: 25ZZ20020446)

Supported and funded by Shandong Provincial Social Science Planning Research Project, titled A Study on the Translation and International Communication of Lv Opera from the Perspective of Communication Studies.

Conflicts of interest

All authors declare that they have no conflicts of interest.

References

Ajitha, P., Sivasangari, A., Rajkumar, R.I. and Poonguzhali, S. (2020) 'Design of text sentiment analysis tool using feature extraction based on fusing machine learning algorithms', *Journal of Intelligent and Fuzzy Systems*, Vol. 40, No. 1, pp.1–9.

Bai, L. (2021) 'Intelligent body behavior feature extraction based on convolution neural network in patients with craniocerebral injury', *Mathematical Biosciences and Engineering*, Vol. 18, No. 4, pp.3781–3789.

Boztas, G. and Tuncer, T. (2021) 'A fault classification method using dynamic centered one-dimensional local angular binary pattern for a PMSM and drive system', *Neural Computing and Applications*, Vol. 34, No. 3, pp.1981–1992.

Chai, Y. (2021) 'Design and implementation of English intelligent communication platform based on similarity algorithm', *Complexity*, Vol. 2021, No. 2, pp.1–10.

Dong, D. (2024) 'Research into a risk assessment model for online public opinions based on big data: random forest and logistic model', *International Journal of Data Science*, Vol. 9, No. 1, pp.19–34, https://doi.org/10.1504/IJDS.2024.135966

Elouariachi, I., Benouini, R., Zenkouar, K. and Zarghili, A. (2020) 'Robust hand gesture recognition system based on a new set of quaternion tchebichef moment invariants', *Pattern Analysis and Applications*, Vol. 23, No. 3, pp.1337–1353.

Fang, Z. and Wang, S. (2024) 'Boosting financial market prediction accuracy with deep learning and big data: introducing the CCL model', *Journal of Organizational and End User Computing*, Vol. 36, No. 1, pp.1–25.

Fei, X. and Tian, G. (2020) 'Attendance automatic recognition and learning behavior of web-based course attendance based on machine learning algorithms', *Journal of Intelligent and Fuzzy Systems*, Vol. 39, No. 2, pp.1–9.

Gan, W., Wang, H., Gu, H., Duan, Y. and Xu, Z. (2021) 'Automatic segmentation of lung tumors on ct images based on a 2d and 3d hybrid convolutional neural network', *The British Journal of Radiology*, Vol. 94, No. 1126, p.20210038.

Hou, Q., Li, C., Kang, M. and Zhao, X. (2020) 'Intelligent model for speech recognition based on SVM: a case study on English language', *Journal of Intelligent and Fuzzy Systems*, Vol. 40, No. 7, pp.1–11.

Jiang, L., Nie, W., Zhu, J., Gao, X. and Lei, B. (2022) 'Lightweight object detection network model suitable for indoor mobile robots', *Journal of Mechanical Science and Technology*, Vol. 36, No. 2, pp.907–920.

Kim, E.K., Jin, Y.K. and Kim, S. (2020) 'Position compensation method for vision-based AGV using feature extraction method based on color space', *Journal of Korean Institute of Intelligent Systems*, Vol. 30, No. 4, pp.251–257.

Li, B. (2020) 'Study on the intelligent selection model of fuzzy semantic optimal solution in the process of translation using English corpus', *Wireless Communications and Mobile Computing*, Vol. 2020, No. 5, pp.1–7.

Li, J., Tu, Y. and Lu, B. (2021) 'Design of intelligent early warning robot system for warehouse autonomous patrol based on scale estimation algorithm of probability theory', *IOP Conference Series: Materials Science and Engineering*, Vol. 1179, No. 1, p.012002(13pp).

Li, X. and Geng, S. (2020) 'Research on sports retrieval recognition of action based on feature extraction and SVM classification algorithm', *Journal of Intelligent and Fuzzy Systems*, Vol. 39, No. 4, pp.5797–5808.

Li, Z., Fan, J., Ren, Y. and Tang, L. (2020) 'A novel feature extraction approach based on neighborhood rough set and PCA for migraine RS-FMRI', *Journal of Intelligent and Fuzzy Systems*, Vol. 38, No. 6, pp.1–11.

Luckhoo, B.F. and Peer, A.A.I. (2023) 'A DEA-WEI method for ranking universities in the presence of imprecise data', *Journal of Data Science*, Vol. 8, No. 3, pp.211–239, https://doi.org/10.1504/Ijds.2023.132285

Meghanathan, N. (2024) 'Assortativity analysis of complex real-world networks using the principal components of the centrality metrics', *International Journal of Data Science*, Vol. 9, No. 1, pp.79–97, https://doi.org/10.1504/IJDS.2024.135945

Meng, F., Jiang, S., Moses, K. and Wei, J. (2023) 'Propaganda information of internet celebrity influence: young adult purchase intention by big data analysis', *Journal of Organizational and End User Computing*, Vol. 35, No. 1, pp.1–18.

Mintorini, E. and Mahmud, W. (2020) 'Rabbit type classification using multi-svm based on feature extraction', *Journal of Applied Intelligent System*, Vol. 4, No. 2, pp.96–103.

Ouariachi, I.E., Benouini, R., Zenkouar, K., Zarghili, A. and Fadili, H.E. (2022) 'RGB-D feature extraction method for hand gesture recognition based on a new fast and accurate multi-channel Cartesian Jacobi moment invariants', *Multimedia Tools and Applications*, Vol. 81, No. 9, pp.12725–12757.

Pan, J., Zhang, J., Wang, F., Liu, W. and Li, Y. (2021) 'Automatic sleep staging based on EEG-EOG signals for depression detection', *Intelligent Automation and Soft Computing*, Vol. 28, No. 1, pp.53–71.

Thotapalli, P.K., Kumar, C. and Reddy, B. (2021) 'Feature extraction of moving objects using background subtraction technique for robotic applications', *International Journal of Intelligent Robotics and Applications*, Vol. 5, No. 2, pp.1–14.

Tian, H. (2023) 'Research on the driving mechanism of business model innovation of startups based on big data analysis in the context of digital economy', *International Journal of Data Science*, Vol. 8, No. 3, pp.195–210, https://doi.org/10.1504/IJDS.2023.132294

Wang, L., Sun, J. and Li, T. (2020) 'Intelligent sports feature recognition system based on texture feature extraction and SVM parameter selection', *Journal of Intelligent and Fuzzy Systems*, Vol. 39, No. 4, pp.1–12.

Wang, P., Cai, H.G. and Wang, L.K. (2020) 'Design of intelligent English translation algorithms based on a fuzzy semantic network', *Intelligent Automation and Soft Computing*, Vol. 26, No. 3, pp.519–529.

Wang, Q., Zong, B., Lin, Y., Li, Z. and Luo, X. (2023) 'The application of big data and artificial intelligence technology in enterprise information security management and risk assessment', *Journal of Organizational and End User Computing*, Vol. 35, No. 1, pp.1–15.

Wang, Z., Cheng, Z., Guan, C. and Han, H. (2021) 'Intelligent information extraction algorithm of agricultural text based on machine learning method', *Journal of Physics: Conference Series*, Vol. 1952, No. 2, p.022073(5pp).

Yang, H. and Yang, Y. (2020) 'Design of English translation computer intelligent scoring system based on natural language processing', *Journal of Physics: Conference Series*, Vol. 1648, No. 2, p.022084(5pp).

Yao, D., Liu, H., Yang, J. and Zhang, J. (2021) 'Implementation of a novel algorithm of wheelset and axle box concurrent fault identification based on an efficient neural network with the attention mechanism', *Journal of Intelligent Manufacturing*, Vol. 32, No. 3, pp.729–743.

Zhang, Z. (2024) 'Deep analysis of time series data for smart grid startup strategies: a transformer-LSTM-PSO model approach', *Journal of Management Science and Operations*, Vol. 2, No. 3, pp.16–43, DOI: https://doi.org/10.30210/JMSO.202402.008

Zhou, K. and Zhang, X. (2020) 'Design of outdoor fire intelligent alarm system based on image recognition', *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 34, No. 07, pp.2765–2766.