International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 6682-6695

https://doi.org/10.70517/ijhsa463574

# Improving Semantic Expression Accuracy in English Translation Teaching Based on Multimodal Learning Models

**Pei Wang**[1,*]
[1] School of Foreign Languages, University of Sanya, Sanya, Hainan, 572000, China
Corresponding authors: (e-mail: wangpei123321@126.com).

**Abstract** English translation teaching faces the challenge of insufficient semantic expression precision, and traditional teaching methods are difficult to capture cross-linguistic semantic nuances. In this study, a deep semantic space model is constructed based on the three principles of relevance, consensus and complementarity, which extracts image features through techniques such as transfer learning, feature adaptation and convolutional neural network, and utilizes bag-of-words model and recurrent neural network for text semantic learning. The experiments validate the model performance on two datasets, TGIF and MSVD, and the results show that on the TGIF dataset, the deep semantic spatial model (DSS) proposed in this paper achieves the R@1, R@5, and R@10 metrics of 9.97, 25.97, and 34.53, respectively, in the video dimension of text retrieval; and the corresponding metrics in the video retrieval of the video dimension are, in order, 15.06, 30.73, 41.22, significantly better than the comparison algorithm. Teaching application experiments show that the translation scores of students in the experimental class with the model-assisted teaching (12.14±1.76) are significantly higher than those of the control class with the traditional teaching method (9.73±2.46), and the difference is statistically significant ($P<0.001$). The study shows that the deep semantic space model based on multimodal learning can effectively improve the semantic expression accuracy in English translation teaching, which provides new technical support and methodological reference for the reform of English translation teaching.

**Index Terms** multimodal learning, English translation teaching, deep semantic space, semantic expression precision, transfer learning, recurrent neural network

## I. Introduction

With the further deepening of economic globalization and political integration, the international community interacts frequently, the social demand for translators is increasing, and the cultivation of qualified translators is becoming one of the important tasks of schools [1]-[3]. In the global language service market, international business, academic communication, scientific research, education and teaching, etc., all need strong translation ability, but in the actual service, the semantic error rate of professional translators has always existed, which reduces the quality of service [4]-[7]. Therefore, schools in various regions have begun to set English translation courses as compulsory or elective courses.

In a broad sense, translation is an indispensable and organic part of the whole English teaching process including listening, speaking, reading and writing. The cultivation of translation ability in English teaching is directly or indirectly closely related to other non-translation teaching contents, which is of considerable significance to the whole English teaching [8], [9]. Once again, from the point of view of the existing English textbooks' writing instructions and writing contents, translation has become an indispensable content in university English teaching. However, there are still many dilemmas in English translation teaching, which hinder the improvement of English teaching quality. In some specialized academic translations, the phenomenon of mistranslation of some traditional literature, legal regulations, and medical texts significantly exists, especially in literature, the semantic loss of literary metaphors and imagery rhetoric is serious, the contextual suitability of legal and medical contexts is low, and the semantic correlation between vocabulary leads to the translation presenting ambiguities and cultural differences, which leads to the impaired precision of final translation semantic expression [10]-[14]. Multimodal learning is a method of learning and reasoning using many different types of data (e.g., images, text, audio, etc.) [15]. In school education, multimodal learning has a wide range of applications. With the introduction of multimodal learning, the teaching mode can become richer and more diverse. Teachers can play audio related to the text and show related pictures and videos to help students promote language comprehension from auditory, visual, and kinesthetic perspectives, and improve ambiguous sentence translation accuracy and metaphor comprehension [16]-[18].

Semantic expression, as the core link of the translation process, directly determines the quality and accuracy of translation. In English translation teaching, students generally have the problems of imperfect semantic understanding and imprecise expression, which mainly stems from the significant differences in lexical-semantic relations, cultural backgrounds and expression habits between English and Chinese languages. The semantic relations between English and Chinese are complex and varied, including correspondence, proximity, parallelism, encompassing, intersection, substitution, vacancy and conflict, which bring great challenges to translation learners. Traditional translation teaching methods focus on the explanation of vocabulary, grammar and skills, while the cultivation of semantic expression accuracy is relatively unsystematic and ineffective, and fails to make full use of modern information technology to improve the teaching effect. Under the background of big data era, multimodal information processing technology provides new development opportunities for English translation teaching. By integrating different forms of information such as text, image, audio, etc., multimodal learning is able to build a more comprehensive and three-dimensional semantic cognitive framework, which helps learners to deeply understand and accurately express cross-linguistic semantic connotations. In particular, the development of deep learning technology provides powerful technical support for multimodal representation learning, which makes it possible to establish semantic mapping relationships between heterogeneous data. At present, the application of artificial intelligence technology in English language teaching has shown great potential, such as dynamic information retrieval, real-time translation language control, and speech recognition and semantic error correction, etc. are gradually changing the traditional teaching mode, but how to organically integrate these technologies into the translation teaching system, especially how to utilize the multimodal learning model to improve the semantic expression accuracy, is still a research topic with urgent exploratory value. Current research shows that the teaching method of combining machine translation and human translation has obvious advantages, but the existing machine translation system is still insufficient in semantic grasping, and it is difficult to meet the needs of high-quality translation teaching. Therefore, the development of more intelligent and accurate semantic analysis and expression models is of great significance to promote the reform of English translation teaching.

Based on the basic principles of multimodal learning algorithms, this study constructs a deep semantic space model, aiming at realizing the mapping and interaction of image and text information in a common semantic space by means of a transfer learning scheme and a feature adaptation method. The study adopts deep convolutional neural network to extract image features, combines bag-of-words model and recurrent neural network for text semantic learning, and establishes a framework that can accurately capture and express cross-modal semantic relations. The effectiveness of the model in improving the semantic expression accuracy of English translation teaching is verified through performance tests on standard datasets and practical teaching application experiments. The study combines the advantages of machine translation and human translation, explores the teaching modes of comparative analysis method, practical teaching method and classroom rehearsal method, provides technical support and methodological guidance for university English translation teaching, and helps to cultivate students' cross-language semantic comprehension and expression ability.

## II. Proposal of a deep semantic space model

### *II. A. Semantic Expression in English Translation Teaching*

#### II. A. 1) “Semantic symbol model”

In the related theory of translation, semantics is the meaning of language. And the relationship between the meaning of the language is linked by a variety of relationships, to a certain extent, semantics is a more general concept that contains many aspects of the meaning and the logic of the relationship.

The semantic theory of translation in the process of development, the level of semantics and various relationships have been systematized research, to provide certain theoretical basis and theoretical support for semantic interpretation and translation.

Semantic symbols are roughly divided into 3 types:

The first type is the sign referent meaning, which is mainly manifested in the interrelationship between the sign and the physical object signified by the sign. However, there is still a deviation between the object signified by the symbol and the physical object, and the symbol can represent both an act and a process, as well as the nature or concept, which is the relationship between each physical object presented by the symbol and its referent. A symbol can be represented as a physical object in real life, or as a process or phenomenon. The sign referent, on the other hand, expresses the relationship between the sign and the signifier, which is the sign's referential meaning.

The second type is the sign-pragmatic meaning, which is mainly expressed in the interrelationship between the sign and the individual who uses the sign, focusing on the relationship between the individuals involved in the process of using language signs and the language signs they use. This encompasses both the individuals who use the symbols and the subjective attitudes of the individuals who use the symbols. In the process of symbol use, this subjective attitude is fused into the symbols, and then with the help of symbol transfer, the meaning is transferred to the physical object signified by the symbols, and this subjective attitude is referred to as a pragmatic relation. For example, sentences with emotional coloring, the use of rhetorical devices, etc., are pragmatic relations, the sign semantics collectively referred to as the pragmatic meaning of the sign.

The third type is intra-language meaning, in which language symbols are interconnected with each other and are an important part of a language system. There exists a certain connection between any symbol and the same symbol, and the symbols are the internal relations of language in the same symbol system, thus it is called the internal meaning of language symbols.

#### II. A. 2) Lexical semantic relations

From the perspective of comparative research between English and Chinese, the semantic relationship between English and Chinese words includes: correspondence, approximation, parallelism, pregnancy, intersection, substitution, vacancy, and conflict. The differences between English and Chinese meanings include: derogatory praise, width and narrowness, old and new, emotional color, national charm, national characteristics, and pragmatic background". For example, the word opera refers to opera that originated in the West, and Peking Opera in China originated in Beijing. Although there are similarities with it, the difference is large, and the semantic relationship between the two is intersecting. Play in English is basically equivalent to Chinese drama, and the semantic relationship between the two is similar. The English rice can summarize the Chinese "rice, rice, grain, rice". temple can summarize the Chinese "temple, temple, sanctuary, shrine, church", and these semantic relationships belong to the pregnant concept.

### *II. B.Application of Intelligent Technology in English Translation*

#### II. B. 1) 2.2.1 Dynamic information retrieval

College students' English course knowledge learning needs are different, and college English translation teaching based on artificial intelligence should likewise be based on students' individualized learning differences, and do a good job of educational practice and educational design.

Among them, the use of dynamic information retrieval function can be based on the language model, for the analysis of English semantic information content, to help students according to their own learning needs, personalized English learning content recommendation. And to develop a systematic learning program for students, so that students can gradually make up for their shortcomings in English learning according to the rhythm of the teacher's English course teaching.

Compared with the traditional teaching mode, the university English translation teaching based on artificial intelligence has stronger educational flexibility. Teachers can use the flexible teaching mechanism to carry out educational guidance to further meet the needs of students for English knowledge learning in various aspects. Let English translation teaching can become an effective way to improve the quality of students' knowledge learning and enrich their knowledge reserves, and realize the high quality of university English translation teaching [19].

#### II. B. 2) Real-time translation of language cross-references

University English course teaching, usually in accordance with the different teaching content of each unit of the course, targeted educational tasks to ensure the effectiveness of the teaching practice of each unit of the course.

The application of artificial intelligence in university English translation teaching can provide teachers and students with visual data analysis and content presentation educational support, so that students can be more three-dimensional knowledge learning.

#### II. B. 3) Speech recognition and semantic error correction

The use of speech recognition to correct problems, help students to further develop the summary of learning problems, and provide teachers with a wealth of educational content, information reference, is an important application of artificial intelligence in university English translation teaching.

At the same time, teachers can also use the natural language processing model to evaluate students' English knowledge learning at all stages, so that students' English translation knowledge learning becomes an important way to test their learning results. Therefore, the use of artificial intelligence speech recognition and semantic error correction for interactive teaching can strengthen students' oral expression ability through English translation teaching, facilitate the deepening of the cultivation of students' English language logical thinking, and enhance the timeliness and relevance of students' English knowledge learning.

### *II. C.Multimodal learning algorithm*

Multimodal representation learning is the core problem of multimodal learning algorithms. In multimodal representation learning, three common basic principles are generally followed: the principle of relevance, the principle of consensus and the principle of complementarity.

#### II. C. 1) Principle of relevance

The goal of the correlation principle is to maximize data correlation across modalities. Suppose that given a pair of data $X = [x_1, \dots, x_n]$ and $Y = [y_1, \dots, y_n]$, the principle of relevance consists in finding the mapping matrices $\omega_x$ and $\omega_y$ such that the corresponding examples in the two datasets are maximally correlated in the mapping space. To wit:

$$\rho = \max_{\omega_x,\omega_y} corr\left(\omega_x^T X, \omega_y^T Y\right) \tag{1}$$

where $corr(\cdot)$ represents the correlation function between $\omega_x^T X$ and $\omega_y^T Y$ .

A typical algorithm for the correlation principle is the typical correlation analysis (CCA) and its various extended transformations.CCA computes the shared embedding of two sets of variables by maximizing the correlation between them and has the ability to model the relationship of the variables very well. As mentioned earlier, CCA attempts to find the linear mappings $\omega_x$ and $\omega_y$ that maximize the correlation between the corresponding examples in the two datasets in the projective space. The correlation coefficient is given by the following equation:

$$\rho = corr(\omega_x^T X, \omega_y^T Y) = \frac{\omega_x^T C_{xy} \omega_y}{\sqrt{(\omega_x^T C_{xx} \omega_x)(\omega_y^T C_{yy} \omega_y)}} \tag{2}$$

The covariance matrix $C_{xy}$ is defined as follows:

$$C_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)^T \tag{3}$$

where $\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the two modal data mean values. The $C_{xx}$ and $C_{yy}$ definitions are similar to the above.

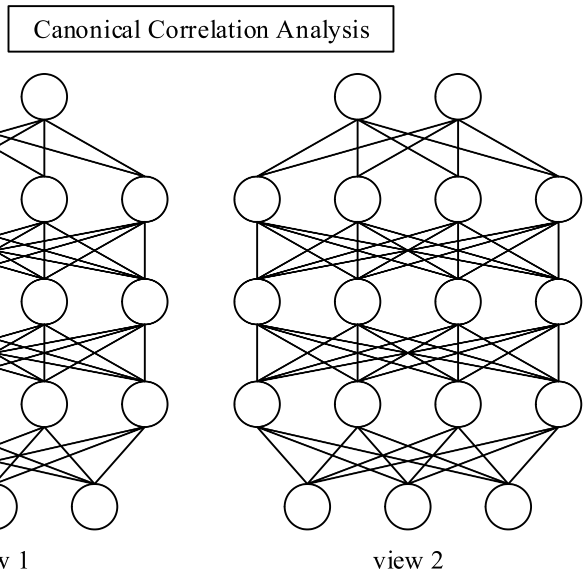


Figure 1: Deep CCA structure diagram

Since the correlation coefficient $\rho$ is invariant to scaling for $\omega_x$ and $\omega_y$ , it is equated to a constrained optimization problem:

$$\begin{gathered}\max_{\omega_x,\omega_y} \omega_x^T C_{xy}\omega_y \\ s.t.\ \omega_x^T C_{xx}\omega_x = 1,\ \omega_y^T C_{yy}\omega_y = 1\end{gathered} \tag{4}$$

By means of the Lagrangian dyad, it can be shown that the solution of Eq. (4) is equivalent to solving a pair of generalized eigenvalue problems:

$$\begin{gathered}C_{xy}C_{yy}^{-1}C_{yx}\omega_x = \lambda^2 C_{xx}\omega_x \\ C_{yx}C_{xx}^{-1}C_{xy}\omega_y = \lambda^2 C_{yy}\omega_y\end{gathered} \tag{5}$$

With the rise and popularity of deep learning, various neural networks have been applied to multimodal data to extract higher-level semantic features, and Deep CCA, a deep typical correlation analysis combined with CCA, has emerged. The block diagram of Deep CCA is shown in Fig. 1.Deep CCA learns the nonlinear mapping representation of different modal data {X,Y} through multi-layer stacked network layers.

$$\begin{gathered}f_x(x) = h_{W_x,b_x}(x) \\ f_y(y) = h_{W_y,b_y}(y)\end{gathered} \tag{6}$$

$W_x$, $W_y$, $b_x$, $b_y$ are the model parameters connecting the layers of the network.The goal of Deep CCA is to jointly learn to optimize the model parameters of the two modal data so that the correlation $corr\left(f_x(X), f_y(Y)\right)$ of the relevant data is as large as possible. Suppose that $\theta_x$ is used to denote all model parameters $b_x$ for the first modal data and $\theta_y$ is used to denote all model parameters $\left(W_y, b_y\right)$ for the second modal data, then:

$$(\theta_x^*, \theta_y^*) = \arg\max_{(\theta_x,\theta_y)} corr\left(f_x(X;\theta_x), f_y(Y;\theta_y)\right) \tag{7}$$

The model training method of Deep CCA is basically the same as other deep neural network training methods. The training error is calculated based on the training data and the predefined objective loss function, and this error is iterated using the back-propagation algorithm for updating the parameters, and the optimal parameter solution of the whole model is finally obtained.

### II. C. 2) Principle of consensus

The purpose of the consensus principle is to maximize the search for consistency in the representations of different modal data. Suppose there are two different modal data $X$ and $Y$, and the feature representations $f(x;W_f)$ and $g(y;W_g)$ of the different modal data are learned by some algorithms or using some neural networks. Here $\theta = \{W_f, W_g\}$ denotes the corresponding model parameters. Assuming that $x_i$ and $y_i$ denote the $i$ th paired data pair, the general measure of the consensus principle is as follows:

$$\min_{\theta} \left\| f(x_i;W_f) - g(y_i;W_g) \right\|_2^2 \tag{8}$$

There are many examples of multimodal deep representation learning based on the consensus principle, and a typical application is the Corresponding Self-Encoder (Corr-AE).Corr-AE starts by constructing two unimodal deep selfencoders, where the two sub-networks are connected by means of a predefined similarity metric on a specific coding layer. Assume that $f(x;W_f)$ and $g(y;W_g)$ are used to denote the feature representations of the two modal data $\{X,Y\}$ obtained through the respective self-encoders, where $\theta = \{W_f, W_g\}$ denotes the corresponding model parameters. The $L_2$ regularity is chosen to measure the similarity of the $i$ th pair of modal data $x_i$ and $y_i$ :

$$C(x_i, y_i;\theta) = \left\| f(x_i;W_f) - g(y_i;W_g) \right\|_2^2 \tag{9}$$

The total objective loss function for the entire Corr-AE model is designed as follows:

$$\begin{gathered}L(x_i, y_i;\theta) = (1-\alpha)(L_I(x_i, y_i;\theta) + L_T(x_i, y_i;\theta)) \\ +\alpha L_C(x_i, y_i;\theta)\end{gathered} \tag{10}$$

In the formula:

$$\begin{aligned} L_I(x_i, y_i; \theta) &= \|x_i - \hat{x}_i\|_2^2 \\ L_T(x_i, y_i; \theta) &= \|y_i - \hat{y}_i\|_2^2 \\ L_C(x_i, y_i; \theta) &= C(x_i, y_i; \theta) \end{aligned} \tag{11}$$

$L_I$ and $L_T$ denote the reconstruction loss of the two modal data through the self-encoder, and $L_C$ denotes the correlation loss, respectively. $\alpha$ is used to adjust the weighting ratio of the two loss functions, and $\hat{x}_i$ and $\hat{y}_i$ denote the reconstructed data of the two modal data $x_i$ and $y_i$, respectively.

### II. C. 3) The principle of complementarity

A typical applied model based on the principle of complementarity is the Deep Boltzmann Machine (DBM). The Restricted Boltzmann Machine (RBM) is an undirected graph model that can learn the distribution of training data. It generally contains a visible layer $v \in \{0,1\}^{d_v}$ and a hidden layer $h \in \{0,1\}^{d_h}$, where the neurons in the same layer are independent of each other and are not connected to each other, and the neurons in different layers are connected to each other.

Since RBM is an energy-based probability distribution model, its goal is to minimize the energy function $E:\{0,1\}^{d_v+d_h} \to R$:

$$E(v,h;\theta) = -\sum_{i=1}^{d_v}\sum_{j=1}^{d_h} v_i W_{ij} h_j - \sum_{i=1}^{d_v} b_i v_i - \sum_{j=1}^{d_h} a_j h_j \tag{12}$$

where $\theta = \{a, b, W\}$ is a parameter of the model. The joint distribution of visible and hidden cells is defined as follows:

$$P(v,h;\theta) = \frac{1}{Z(\theta)} \exp\left(-E(v,h;\theta)\right) \tag{13}$$

A deep Boltzmann machine is a generative network. It consists of a visible layer $v \in \{0,1\}^{d_v}$ and several hidden layers $h^{(1)} \in \{0,1\}^{d_{h1}}, h^{(2)} \in \{0,1\}^{d_{h2}}, \cdots, h^{(L)} \in \{0,1\}^{d_{hL}}$. Taking the example of a DBM with two hidden layers, the energy of the joint representation $\{v,h\}$ is defined as follows:

$$E(v,h;\theta) = -v^T W^{(1)} h^{(1)} - h^{(1)} W^{(2)} h^{(2)} \tag{14}$$

Deep Boltzmann machines can be easily extended to multimodal representation learning. Assuming that two modal data $v_m$ and $v_t$ are modeled, each modal data is modeled and represented with a separate two-layer DBM, and then an additional network layer is added on top of them as a shared representation layer. The distribution of $v_m$ is given by the following equation:

$$\begin{aligned} P(v_m;\theta) &= \sum_{h^{(1)},h^{(2)}} P\left(v_m, h^{(1)}, h^{(2)}; \theta\right) \\ &= \frac{1}{Z(\theta)} \sum_{h^{(1)},h^{(2)}} \exp\left( -\sum_{i=1}^{d_{vm}} \frac{(v_{mi} - b_i)^2}{2\sigma_i^2} \right. \\ &\quad \left. + \sum_{i=1}^{d_{vm}} \sum_{j=1}^{d_{h1}} \frac{v_{mi}}{\sigma_{ij}} W_{ij}^{(1)} + \sum_{j=1}^{d_{h1}} \sum_{i=1}^{d_{h2}} h_j^{(1)} W_{ji}^{(2)} h_i^{(2)} \right) \end{aligned} \tag{15}$$

The distribution of $v_t$ is defined similarly to the above equation. The joint distribution of multimodal inputs can be written as follows:

$$\begin{aligned} P\left(v_m, v_t; \theta\right) &= \sum_{h_m^{(1)}, h_m^{(2)}, h^{(3)}} P\left(h_m^{(1)}, h_m^{(2)}, h^{(3)}\right) \\ &\left( \sum_{h_m^{(1)}} P\left(v_m, h_m^{(1)}, h_m^{(2)}\right) \right) \left( \sum_{h_t^{(1)}} P\left(v_t, h_t^{(1)}, h_t^{(2)}\right) \right) \end{aligned} \tag{16}$$

### *II. D.Deep semantic space modeling*

Traditional cross-modal retrieval methods have limitations, including the use of low-level features of complex images, making it difficult for these methods to exploit high-level semantic associations between different modal data. There is a lack of ability to handle data in a single modality that does not occur in pairs.

In order to compensate for the shortcomings of the traditional cross-modal retrieval methods mentioned above, this paper proposes a cross-modal retrieval framework based on deep semantic learning to learn the common semantic space of different modal data. Algorithm performance tests are carried out to verify the feasibility of the algorithm, and the algorithm is used in English translation teaching as an auxiliary tool to analyze its use in English translation teaching.

(1) Migration learning program

For image semantics, the concept of transfer learning is introduced to migrate the a priori semantic knowledge of ImageNet in the source domain to the target domain. The source domain ImageNet is a large image library for target recognition in computer vision. The target domain is the Wiki-Flickr event dataset of this paper. Next, the deep convolutional neural network VGGNet is utilized for image feature extraction, and finally the features of the image are converted into high-level deep semantic information by the nonlinear transformation of the fully connected neural network. A scheme of migration learning with layer-by-layer fine-tuning of the pre-trained model is used in the experiments.

(2) Feature adaptation method

In this paper, we adopt the method of domain adaptation to solve the problem of domain contradiction by minimizing the maximum mean difference (MMD) between the data in the source domain as well as the data in the target domain of the same modality, so that the migration model can better match the distribution of the data in the target domain. Defining the maximum mean difference between the image from the source domain $\{i^s\}$ with distribution $c$ and the image from the target domain $\left\{i^t\right\}$ with distribution $d$ as $d_k(c,d)$, the square of MMD in the regenerated kernel Hilbert space $H_k$ is The formula is defined as:

$$d_k^2(c,d) = \left\| E_c[\phi(i^s)] - E_d[\phi(i^t)] \right\|_{H_k}^2 \tag{17}$$

where $\phi$ denotes the representation of a particular layer in a deep neural network, and let the average embedding in the regenerated kernel Hilbert space $H_k$ be $u_k(c)$ for images with distribution $c$, and $E_{X\sim c} f(x) = \left\langle f(x), u_k(c) \right\rangle_{H_k}^2$ be satisfied for all $f \in H_k$. The loss function for migration is:

$$Loss_{Transfer} = \sum_{l=l_6}^{l_7} d_k^2\left(I^s, I^t\right) \tag{18}$$

where $d_k^2(I^s, I^t)$ the migration loss of a layer in the neural network, and $l_6$ and $l_7$ denote the different fully connected layers, respectively. The contradiction between the source domain as well as the target domain can be effectively reduced by the migration loss function in the above equation.

(3) Convolutional neural network for image feature extraction

For image feature extraction, this paper utilizes the method of deep convolutional neural network VGGNet to automatically learn the characterization of the image. Compared to traditional image feature extraction methods, end-to-end learning can be performed. Better performance gains can be obtained by using deep learning methods with more expressive capabilities than traditional machine learning methods.

(4) Image Semantic Learning

Finally, the extracted image features are input to the fully connected neural network, and the features of the image are converted into high-level deep semantic information using nonlinear transformation. Let the output of the last fully connected layer of VGGNet be $O_I$, and input $O_I$ into the $k$-dimensional softmax activation function for generating the semantic embedding of the image $s_I \in R^k$, where $k$ is the number of categories. The semantic embedding of the image is defined as:

$$\sigma : R^k \to \left\{ z \in R^k \mid z_i \geq 0, \sum\nolimits_{i=1}^{k} z_i = 1 \right\} \tag{19}$$

$$z_j = (O_I)_j \ for \ j = 1,2,\ldots,k \tag{20}$$

$$(S_I)_j = P\left(y = j \mid I\right) = \sigma(z)_j = \frac{e^{z_j}}{\sum_{i=1}^{k} e^{z_k}} \ for \ j = 1,2,\ldots,k \tag{21}$$

where $P\left(y = j \mid I\right)$ denotes the conditional probability of predicting that an image belongs to the $j$th class given an image $I$, $s_I \in R^k$ is the semantic embedding of the image, and the subscript $j$ represents the $j$th element of the vector. Intuitively, the softmax function maps the input $O_I$ into a $k$-dimensional probability vector, and the sum of the elements in the vector is 1.

(5) Text Encoding

For the learning of text semantics, this paper adopts the bag-of-words model in natural language processing to encode the text. Bag-of-words model is a kind of text representation. The process of text semantic learning: text preprocessing, text feature vectorization, text semantic learning.

Preprocessing the document is the preparatory work for information encoding. The process of text preprocessing such as: raw text, segmentation, cleaning, standardization.

After preprocessing, the document can be encoded by text feature vectorization. The methods used in this paper are TF-IDF as well as LDA [20].

TF-IDF algorithm is an algorithm used to evaluate the importance of a word in a document. It considers the frequency of the word in the document (TF) and the frequency of its occurrence in all documents (IDF), and the TF-IDF value is obtained by multiplying the two. The higher the TF-IDF weight of a word in a document, the higher is its importance. The formula involved is as follows:

$$TF_{i,j} = n_{i,j} \Big/ \sum_k n_{k,j} \tag{22}$$

$$IDF_i = \log(N / n_i) \tag{23}$$

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \tag{24}$$

where $n_{i,j}$ is the number of occurrences of word $i$ in document $j$, and $\sum_k n_{k,j}$ is the sum of occurrences of all words in document $j$. $N$ is the total number of documents in the corpus, $n_i$ is the number of documents containing word $i$, and the result is the weight corresponding to word $i$.

(6) Text Semantic Learning

After the extraction of text features as described above, the text is vectorized and then fed into the recurrent neural network to learn the deep semantic space. In this paper, LSTM is used to learn the deep semantic information hidden in the document.

LSTM is an improved recurrent network architecture that uses a gating mechanism consisting of input gates $(i_t)$, forgetting gates $(f_t)$, and output gates $(o_t)$, which help to determine whether the data in the current state should be retained or forgotten from the previous state.The structure of the LSTM is shown in Figure 2.

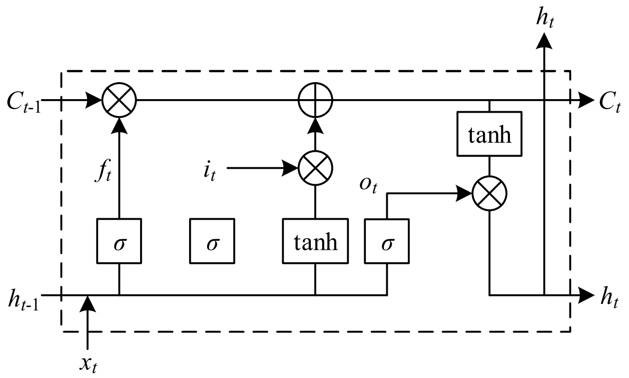


Figure 2: The overall structure of the LSTM

The forgetting gate processes information from the previous memory unit and is calculated as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{25}$$

The input gate determines how much new information is added to cell $c$ and is calculated as follows:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{26}$$

$$c_t = i_t * \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right) + f_t c_{t-1} \tag{27}$$

The output gate determines the output of the current state and is calculated as follows:

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{28}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{29}$$

where $x_t$ is the input signal received at moment $t$, $h_{t-1}$ is the previous output signal of the LSTM at moment $t-1$, $W$ and $b$ are the weights and bias respectively, $\sigma$ is the Sigmoid function, $\tanh$ is the hyperbolic tangent function, $c_t$ is the quality of the neuron's output at moment $t$, and $h_t$ is the final output signal.

(7) Deep Semantic Space Modeling of Interaction

Through the learning of the deep semantic space framework, the originally heterogeneous image and text data are mapped to the isomorphic deep semantic spaces $S_I$ and $S_T$. Given a correlated image-text pair $(s_I, s_T)$, correlation means that the image and text share common labels. Finally, the deep semantic space model is trained end-to-end using relevant image-text pairs as well as irrelevant image-text pairs as a training set, with the goal of minimizing the cosine similarity loss defined by the following equation:

$$L_{\cos}\left((s_I, s_T), y\right) = \begin{cases} 1 - \cos\left(s_I, s_T\right) & if\ y = 1 \\ \max\left(0, \cos(s_I, s_T) - \alpha\right) & if\ y = -1 \end{cases} \tag{30}$$

where $\cos(\cdot)$ denotes the normalized cosine similarity and $\alpha$ is the marginal parameter.

High-level semantic information is embedded in the softmax layer, and the weights of the higher levels in the model are shared. Since the model is also essentially learning to categorize image text, the introduction of semantic loss helps to distinguish the relationship between image text pairs. The final loss function of the model is defined as:

$$L(s_I, s_T, c_I, c_T, y) = L_{cos}((s_I, s_T), y) + \lambda L_{reg}(s_I, s_T, c_I, c_T) \tag{31}$$

where $L_{reg}(s_I, s_T, c_I, c_T)$ is the semantic loss, $\lambda$ is the regularization parameter, and $c_I, c_T$ are the semantic category labels of the image and text, respectively. When $(s_I, s_T)$ is a positive instance, $c_I$ and $c_T$ are identical.

## III. Translation applications of deep semantic space models

### III. A. Model performance

Experimental configuration: the algorithmic models were all implemented using the Pytorch framework, the operating system was Ubuntu 16.04, and all experiments were conducted on an NVIDIA TITAN XP graphics card.

In order to better verify the validity of the models, experiments were performed on the TGIF and MSVD datasets, respectively.

(1) TGIF dataset

In order to verify the effectiveness of the deep semantic space framework (DSS) proposed in this paper on the TGIF dataset, comparison experiments are conducted on the TGIF dataset.

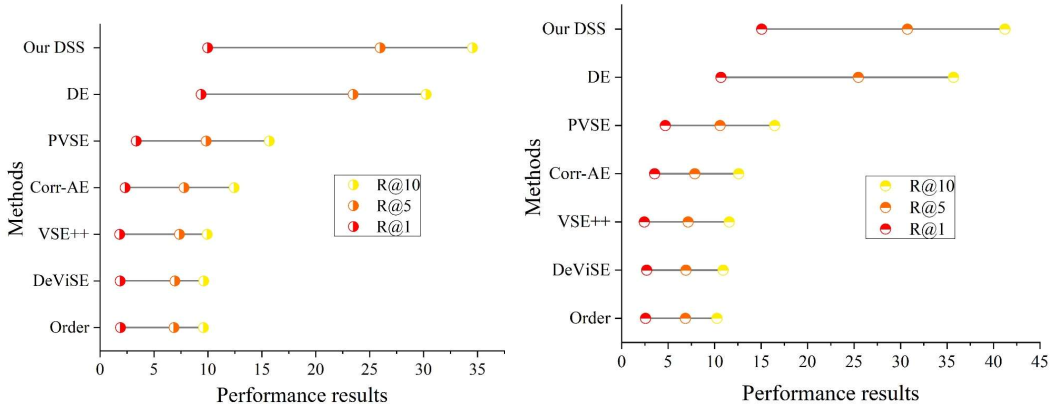
In the TGIF dataset, a GIF has only one text description corresponding to it. While a GIF in the test set has three text descriptions corresponding to it, all compared with some difficulty.

The comparison algorithms chosen include Order, DeViSE, VSE++, Corr-AE, PVSE, DE.

The evaluation metrics used in comparison with the comparison algorithms include R@1, R@5, and R@10.

The performance difference between the model of this paper and the comparison algorithm is shown in Fig. 3. Figures (a) and (b) show the scores of the metrics in the text retrieval video and video retrieval text dimensions, respectively. From the experimental results, it is obvious that DSS with DE outperforms other algorithms in all metrics, whether it is using text to retrieve video or video to retrieve text. And the DSS model scores 9.97, 25.97, 34.53 for all the metrics in the video dimension of text retrieval in TGIF dataset and 15.06, 30.73, 41.22 for all the metrics in the text dimension of video retrieval in that order.

From the above analysis of the experimental results, it can be concluded that the enhancement of DSS model over DE on TGIF dataset is very significant.



(a) Text-Video Retrieval (b) Video-Text Retrieval

Figure 3: The performance difference of the model and the comparison algorithm

(2) MSVD dataset

To further validate the performance of the model, experiments were also done on the MSVD dataset.

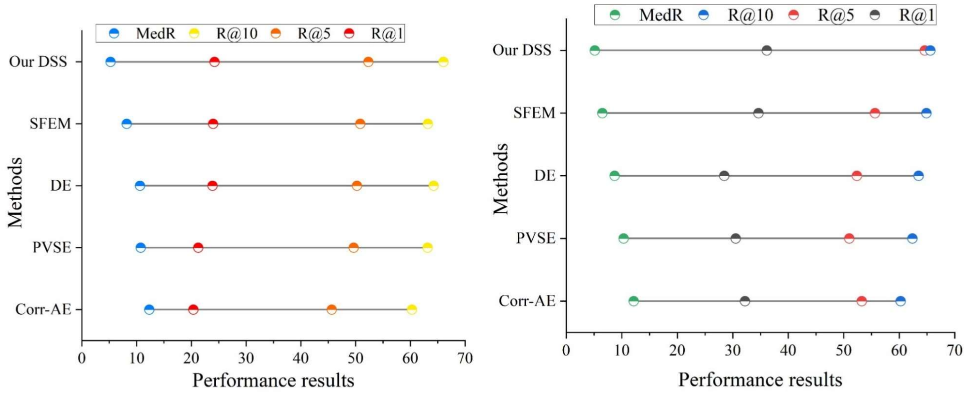
Since many of the comparison algorithms on the MSVD dataset sample text descriptions indeterminately, a sampling approach was used for the experiment in order to fairly compare the performance differences between the algorithms. The comparison algorithms used in this experiment were Corr-AE, PVSE, DE and SFEM.

The evaluation metrics used on this dataset include R@1, R@5, R@10, and MedR.

The difference in performance of different algorithms on MSVD is shown in Fig. 4, Fig. (a) and Fig. (b) show the metrics scores in the text-retrieved video and video-retrieved text dimensions, respectively.

On the R@1 metric, the DSS model outperforms the well-performing SFEM method by 0.27 vs. 1.51, respectively. On the R@5 metric, it outperforms the SFEM method by 1.49 and 8.95, respectively.

Compared with the DSS model, the difference between the DE algorithm and the DS R@10 S model is 1.79 and 2.11, respectively. The MedR values of the DSS model in the text retrieval video dimension and the video retrieval text dimension were 5.24 and 5.12, respectively.



(a) Text-Video Retrieval (b) Video-Text Retrieval

Figure 4: Different algorithms performed on MSVD performance

## III. B. Model application

### III. B. 1) Mode of application

The teaching method of combining machine translation and human translation is a technology-supported and practice-led translation education method, which can combine the advantages of machine translation and human translation to provide students with a more flexible and personalized translation practice environment, so as to improve students' translation ability and translation practice level.

Machine translation can be used to assist students' translation learning, especially when a large number of terms, words and phrases are involved. With the results of machine translation, students can understand the meaning and grammatical structure of the translated text faster, which helps to improve translation speed and accuracy. At the same time, machine translation can also help students expand their vocabulary and familiarize themselves with the expression habits of different languages, thus improving cross-cultural communication skills. The deep semantic space model proposed in this paper is used to enhance semantic learning in English translation and improve English semantic analysis ability.

In the process of translation teaching, combining machine translation and human translation can effectively improve students' translation ability and translation quality. It can be realized through the following methods in teaching practice:

Comparative analysis method: the results of machine translation are compared and analyzed with the results of human translation, so that students can understand the differences and advantages and disadvantages of machine translation and human translation. Students can better understand the influence of language and cultural background on translation and improve the quality of translation by analyzing the results of comparison.

Practical Teaching Method: Through the way of practical teaching, students can experience the process of machine translation and human translation. Students can use machine translation to do preliminary translation, and then use human translation to make corrections and embellishments, and finally achieve better translation results.

Classroom Exercise Method: In the classroom, students are allowed to practice machine translation and human translation in small groups. Students can choose to translate using machine translation or human translation independently according to the knowledge and skills they have learned, and discuss and share their experiences with each other in the group. Through the teaching method of combining machine translation and human translation, students can better master different translation skills, such as vocabulary translation, grammar translation, language style translation, etc., and at the same time, it can also enable students to better understand the differences between different languages and improve the ability of cross-cultural communication. In practice, through the use of modern technological means such as translation software, platforms and tools, students can better apply translation technology and improve the efficiency and quality of translation. And it can provide students with a more comprehensive, in-depth and practically meaningful translation education experience, as well as better meet the social demand for high-level talents, and play an important role in improving students' translation ability and translation practice.

### III. B. 2) Pedagogical applications

Spss20.0 statistical software was used to analyze and compare the scores of students' entrance English exams provided by the Recruitment and Placement Office of a university. Two classes that do not have significant difference in English language level were selected as experimental and control classes among the 24 university English teaching classes in the College of Architecture. There were 42 students in the experimental class and 45 students in the control class.

Pre-test: the test subjects were the students in the experimental and control classes. Test instrument: the real translation questions of the December 2023 Grade 4 exam. Scoring criteria: according to the translation scoring criteria of the University English Grade 4 Examination, assessed by the members of the subject group. Statistical tools: Spss20.0 statistical software to analyze the experimental data.

Without changing the existing university English credits and the total amount of class hours, the experimental class and the control class were taught translation with different contents and teaching methods for 1 credit hour/week, and the experimental cycle was one semester.

Experimental class: in addition to regular college English teaching using conventional textbooks (3 credit hours/week), specialized college English translation teaching was also conducted, with the main teaching content being mechanical translation based on deep semantic space model selected by the members of the group with reference to existing translation textbooks of various kinds to assist learning, as well as the task training module for students to perform translation drills. The teaching methodology adopts the model of translation training for translation teaching. The specific arrangement of 4 classes per week is as follows: 2 intensive reading classes, 1 translation class and 1 listening class per week, and the translation part of the final examination accounts for 20%.

Control Class: In addition to using the regular textbook for regular college English teaching (3 hours/week), college English translation teaching is also conducted, but the teaching content is the translation practice questions in the current textbook and the accompanying exercise book, and the teacher corrects, explains, and corrects the translations of the students after they finish the translation assignments. The specific arrangement of four classes per week was three intensive reading classes (with translation teaching included) and one listening class per week, and the translation part of the final examination accounted for 20%.

Post-test: to test the translation level of the experimental class and the control class after one semester of translation study, to see if there is any significant difference between the translation levels of the two classes. Test instrument: the real translation questions of the June 2024 Grade 4 exam.

The results of the pre-test for the experimental and control classes are shown in Figure 5. It can be seen that the students' scores in both the experimental and control classes do not exceed 18 points. The statistical results obtained were 8.93±2.77 and 9.01±3.27 for the students in the experimental and control classes, respectively, and there was no significant difference between the pre-test results of the two classes.

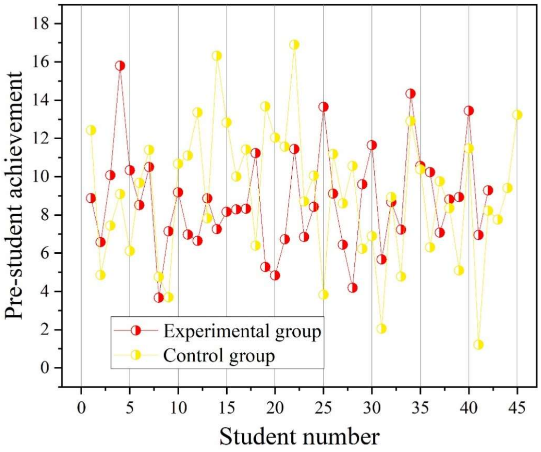


Figure 5: Test results of the experimental class and the comparison class

The post-test scores of the two classes are shown in Figure 6, and it can be seen that the distribution of students' scores in the experimental class is generally higher than the scores of students in the control class. The lowest and highest scores of the students in the experimental class were 7.80 and 16.27, respectively. 12.14±1.76 and 9.73±2.46 were obtained for the experimental and control classes, respectively.

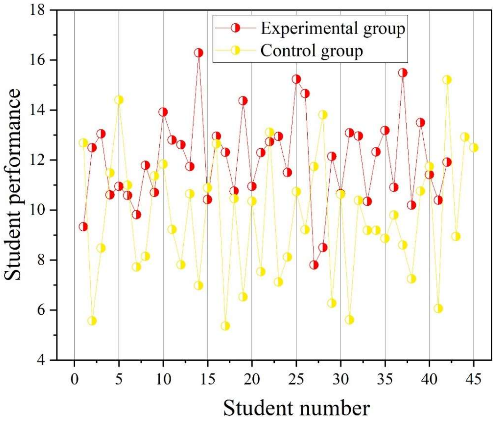


Figure 6: The post-test results of the two classes

Independent samples tests were performed on the posttest scores, and the independent samples tests for the posttest results are shown in Table 1. The T-value was 5.128 and the P-value was 0.000. Therefore, there is a significant difference between the two groups in the posttest translation scores, which means that the experimental group's scores are statistically significantly higher than those of the control group. Thus, it can be shown that the deep semantic spatial model-assisted English translation learning in the experimental group is helpful to students' translation level improvement, and the English translation achievement improvement is higher than the traditional English translation learning. It can be considered that the English semantic translation technology built based on the deep semantic spatial model of multimodal learning model can assist English translation teaching or provide a little inspiration and reference for the reform of university English translation teaching.

Table 1: The results were tested independently

| | The variance equation levene test | | T test of the mean equation | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | Sig. | t | df | Sig. (Double side) | Mean difference | Standard error value | 95% confidence interval of the difference | |
| | | | | | | | | Lower limit | Upper limit |
| The variance is equal | 0.317 | 0.669 | 5.128 | 86 | 0.000 | 1.789 | 0.361 | 0.952 | 2.801 |
| The variance is not equal | | | 5.003 | 80.953 | 0.000 | 1.624 | 0.390 | 0.945 | 2.796 |

## IV. Conclusion

The Deep Semantic Space model shows remarkable results in improving the accuracy of semantic representation in English translation teaching. The model performance test shows that on the TGIF dataset, the DSS model achieves R@1, R@5 and R@10 metrics of 15.06, 30.73 and 41.22 in the video retrieval text dimension respectively, which is superior to the existing algorithms; on the MSVD dataset, the DSS model achieves a MedR value of 5.24 in the text-retrieval video and 5.12 in the video-retrieval text dimensions, which exhibits excellent cross-modal retrieval capability. Teaching application experiments further verified the practical value of the model. After a semester of controlled experiments, the translation scores of the students in the experimental class assisted by the deep semantic space model (12.14±1.76) were significantly higher than those in the control class (9.73±2.46) with traditional teaching methods, and the difference between the two groups was statistically significant (T=5.128, $p<0.001$). This result confirms the promising application of multimodal learning technology in English translation teaching, which effectively enhances students' ability to understand and express the deep semantics of the language by integrating multidimensional information such as text and images. The translation teaching method based on deep semantic spatial modeling integrates the efficiency of machine translation with the accuracy of human translation, which not only improves the efficiency of translation teaching, but also cultivates students' intercultural communication ability. This research provides new ideas for the reform of English translation teaching and also lays the foundation for the wide application of multimodal learning technology in the field of language education.

## Funding

This work was supported by Ministry of Education, Industry-University-Cooperation Project: "Innovative Research on Information - based Teaching of English Translation from the Perspective of International Communication" (Project Number: 230700700212147);

University of Sanya Talent Introduction Project: "Research on Building International Communication Ability of Compound Foreign Language Talents" (Project Number: USYRC24-01).

## References

[1] Wu, Y. (2023). A Study on the Innovative Model of Cultivating Internationalized Composite Translation Talents under the Background of Hainan Free Trade Port. International Journal of New Developments in Education, 5(11).

[2] Brisset, A. (2017). Globalization, translation, and cultural diversity. Translation and Interpreting Studies, 12(2), 253-277.

[3] Hassenteufel, P., & Zeigermann, U. (2021). Translation and translators in policy transfer processes. Handbook of policy transfer, diffusion and circulation, 58-79.

[4] Romadhon, R. (2025). LOST IN TRANSLATION: A CRITICAL ANALYSIS OF ERRORS IN ABSTRACT TRANSLATIONS BY VOCATIONAL ACCOUNTING STUDENTS. Journal of English Language and Culture, 15(1).

[5] Solodkova, I. M., Grigoryeva, E. V., & Ismagilova, L. R. (2021). Translating texts on business and economics: Some pitfalls for perspective translators. EntreLínguas, 7(1), 203-212.

[6] Pudjiati, D., & Fadilah, E. (2017, December). Semantic errors in the translation about actions to defend Islam in 2016 into English. In International Conference on Culture and Language in Southeast Asia (ICCLAS 2017) (pp. 98-102). Atlantis Press.

[7] Islam, S. (2018). Semantic loss in two English translations of surah Ya-sin by two translators (Abdullah Yusuf Ali and Arthur John Arberry). IJLLT, 1(4), 18-34.

[8] Wang, C. (2017, July). Research on Cultivation of Translation Ability and College English Teaching. In 2017 3rd International Conference on Economics, Social Science, Arts, Education and Management Engineering (ESSAEME 2017). Atlantis Press.

[9] Saud, W. I. (2020). The relationship between critical thinking and translation ability of EFL undergraduate students. International Journal of Social Sciences & Educational Studies, 7(3), 19-28.

[10] Wongranu, P. (2017). Errors in translation made by English major students: A study on types and causes. Kasetsart journal of social sciences, 38(2), 117-122.

[11] Afrouz, M. (2024). Overtly-Erroneous Lexico-semantic Errors in Contemporary Resistance Literary Works Translated from Persian into English. Iranian Journal of Translation Studies, 22(85).

[12] Siumarlata, V., Sallata, Y. N., & Tandikombong, M. (2024). A SEMANTIC ANALYSIS OF COMMON ERRORS: EVALUATING THE LIMITS OF ONLINE TRANSLATION APPLICATIONS. English Language Teaching Methodology, 4(3), 540-551.

[13] Saksittanupab, P., & Pattanasorn, C. (2023). An Analysis of Lexical Semantic Errors in English Writing by Thai EFL Learners. Journal of Roi Kaensarn Academi, 8(7), 141-151.

[14] Bracken, J., Degani, T., Eddington, C., & Tokowicz, N. (2017). Translation semantic variability: How semantic relatedness affects learning of translation-ambiguous words. Bilingualism: Language and Cognition, 20(4), 783-794.

[15] Bouchey, B., Castek, J., & Thygeson, J. (2021). Multimodal learning. Innovative learning environments in STEM higher education: Opportunities, challenges, and looking forward, 35-54.

[16] Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. Journal of Computer Assisted Learning, 34(4), 338-349.

[17] Rahmanu, I. W. E. D., & Molnár, G. (2024). Multimodal immersion in English language learning in higher education: A systematic review. Heliyon.

[18] Rohi, M. P., & Nurhayati, L. (2024). Multimodal learning strategies in secondary EFL education: Insights from teachers. Voices of English Language Education Society, 8(2).

[19] Yuhua Wang. (2024). Editorial Expression of Concern: Artificial Intelligence Technologies in College English Translation Teaching.. Journal of psycholinguistic research,53(4),61.

[20] Fei Li,Huishang Li,Xin Dai,Hongjie Ren & Huaiyang Li. (2025). Does Online Public Opinion Regarding Swine Epidemic Diseases Influence Fluctuations in Pork Prices?—An Analysis Based on TVP-VAR and LDA Models. Agriculture,15(7),730-730.