International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 2916-2926

https://doi.org/10.70517/ijhsa463244

Computer semantic analysis and English translation strategies for the philosophical connotations of the character “Body” in “The Analects” from a translation studies perspective

Zijuan Su[1,*]
[1] School of Foreign Languages and Culture, Geely University of China, Chengdu, Sichuan, 641423, China
Corresponding authors: (e-mail: suzijuan@126.com).

Abstract As one of the most important Chinese cultural classics, the study of English translation of The Analects of Confucius is also a hot research topic in the field. Based on statistical machine translation, which is a mainstream method in machine translation, this paper combines two methods of obtaining word vector space representations and word embedding representations, namely, Point Mutual Information (PMI) and Word2vec, to construct a translation model based on word semantic distributions, which provides technical support for the English translation of The Analects of Confucius. The English translation model of this paper is used to integrate and analyze the philosophical connotation of the word “Body” in the Analects, and to explore the acceptability of the English translation results to the readers. The main philosophical connotation of the word "body" in the Analects is to be self-oriented and Tao-oriented. The former is to start all moral practice from self-reflection and cultivation, while the latter is to the individual's "body" beyond the ego and finally integrate into the Confucian order with "benevolence" and "propriety" as the core. The mean syntactic acceptability of Chinese and foreign readers is 3.25 and 3.12 respectively, and the mean main idea acceptability is 3.32 and 3.15 respectively, all of which are in the range of 2.5~3.5, presenting a basically acceptable attitude.

Index Terms point-to-point information, Word2vec, machine translation, Analects of Confucius

I. Introduction

As one of the representatives of traditional Chinese culture, The Analects of Confucius is rich in philosophical connotations [1]. In the ideological perspective of the Analects, the body is not only the material materiality of human beings, but also the cohesion of human nature, and the practice of benevolence and virtue is the process of humane practice in which the body humanizes humanity and carries the Way with the body [2], [3]. In the logical chain of bodily humanism, the body presents three aspects, including the body of moral sense in which virtue springs from within, the body of discipline in which self-restraint is restored, and the body of performance in which virtue is applied to deeds. The body of bodily sensation is not the body of speech, but the body of virtue which is filled with virtue within. The body of discipline is the externalization and embodiment of the body of virtue, the core idea of which is “correcting the body”, and the way of discipline is to return to one's self and “reverse” the body to the way of the body. Virtue is made by “cultivation”, and the Way is carried by “body”; only practicing virtue is the basis of the Way. These three facets are interrelated and roughly present the philosophical logic of the Analects of Confucius in terms of the body. The “body” philosophy of the Analects shows readers a philosophy of ethical behavior, moral beliefs, and self-improvement through education, and the study of its English translation strategy can help the world better understand China [4]-[7].

With the development of technology, the application of English machine translation technology in ancient literature has received much attention [8]. There are countless existing translation model designs on English, which can satisfy users' needs to some extent [9]. However, the same problem exists that most of them are English machine translation models based on lexical disambiguation and semantic role labeling [10], [11]. Therefore, it is of great significance to propose computer semantic models that have the ability to perform semantic expression independently and also have the ability to describe the relationship between words to provide accurate translation services for users in various fields [12], [13].

In order to solve the translation problems faced by the traditional cultural classics represented by the Analects, this paper constructs a translation model based on the semantic distribution of words on the basis of statistical

machine translation. Two methods of obtaining word vector space representation and word embedding representation, namely, point-by-point mutual information (PMI) and Word2vec, are introduced to integrate the sentence and document context information of the phrases to be translated, and estimate the translation probability of the phrases with the target candidate phrases, so as to dynamically make the translation decision according to the context information of the test text. A convolutional neural network model is used to learn the semantic features of the sentence context of the source phrase and the corresponding target candidate phrase, and to obtain the thematic features of the document where the phrase is located. Secondly, the semantic information of the source phrase and the topic information of the document are fused to obtain the fused semantic features of the phrase in the context, and the semantic match between the phrase and the target candidate phrase is calculated to estimate the probability of the phrases being translated into each other in the specified context. Finally, the probability of mutual translation of phrase pairs is added as a new feature into the decoder of the translation system. The translation performance of this translation model is tested, and it is applied to translate the original text of The Analects of Confucius into English, integrating the philosophical connotation of the word “body” in The Analects of Confucius and analyzing the readers' acceptance of the result, so as to explore the practical utility of the translation model in this paper.

II. Translation model based on word semantic distribution

The Analects is a collection of quotations compiled by the disciples and re-transmitted disciples of Confucius, a thinker and educator of the Spring and Autumn Period, who recorded the words and deeds of Confucius and his disciples. Written in the early Warring States period, it is one of the most important cultural classics in China, and has a very high value of cultural dissemination. In the early days, the literary style of the Analects led to the high cost of its manual English translation, and there were big obstacles. With the rapid development of the Internet and artificial intelligence, machine translation-related researches are emerging, and the use of translation models for English translation of ancient books in the literary style represented by the Analects is of great significance for the overseas digital dissemination of China's excellent traditional culture.

In this chapter, a translation model based on the semantic distribution of words will be constructed to utilize the contextual information of sentences and documents in which the source language phrases are located to compute their semantic matches with the target candidates, in order to dynamically adapt to the domain changes of the test text and improve the translation efficiency.

II. A. Overview of statistical machine translation

For different statistical machine translation systems, the training process of the model varies, in this section we take the phrase-based statistical machine translation system as an example to illustrate the training process of the model. The whole process of model training mainly includes: phrase translation table extraction, ordering model training and target language language model training [14].

The extraction and scoring of phrase translation table are mainly completed in the process of lexicalization model training. The phrase translation table mainly consists of three items: phrases on the source language side, phrases on the target language side, and phrase pair feature scores. The phrase pair feature scores usually include features such as: forward translation probability, forward lexicalization probability, reverse translation probability and reverse lexicalization probability.

Target language language modeling training mainly refers to the establishment of a translation language model after the corpus of the target language is trained. The language model can be used to measure the fluency of the output text, which is an integral part of statistical machine translation. At present, the more commonly used statistical language models mainly include context-independent model, n meta-grammatical language model and neural network language model, etc., and the commonly used modeling tools mainly include SRILM and so on.

To summarize, for a given sentence f to be translated, its optimal translation e can be calculated by Equation (1) in a phrase-based statistical machine translation system:

$$\tilde{e} = \arg\max_{e} \left\{ \Pr\left(e|f\right) \right\} \tag{1}$$

where e and f are the target language sentences and source language sentences in statistical machine translation. $\Pr\left(e|f\right)$ is the translation probability, and the specific calculation process is shown in Equation (2):

$$\begin{aligned}\Pr(e|f) &\approx p_{\lambda_1,\lambda_2,\cdots,\lambda_M}(e|f) \\ &= \exp\left[\sum_{m=1}^{M}\lambda_m h_m(e,f)\right] \Big/ \sum_{e'}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(e',f)\right]\end{aligned} \tag{2}$$

The above is the mathematical representation of the phrase-based statistical machine translation method, where h is the M features on the source and target languages, λ is the M weight values corresponding to these features, and the denominator in Equation (2) plays the role of normalization.

II. B.Word Semantic Information Acquisition

From the computational point of view, word semantic information acquisition methods that are more commonly used at present are word vector space representation and word embedding representation. In this section, we introduce two methods to obtain word vector space representation and word embedding representation respectively, i.e., point-by-point mutual information (PMI) and Word2vec.

II. B. 1) Point-by-point mutual information to obtain word vectors

Word vector space representations are easily confused in practical applications of statistical counting, and their values are larger if the contextual words co-occur more often. Therefore, the point-by-point mutual information (PMI) in information theory is usually used to obtain the word vector space representation in practical applications.

Specifically, in the word semantic vector represented by point-by-point mutual information (PMI), the elements in each dimension of the vector reflect the frequency of co-occurrence of the target word with other words in the word list in a predefined context window, and PMI can be calculated by Equation (3):

$$pmi(c_i,t) = \log\frac{p(c_i,t)}{p(c_i)\,p(t)} = \frac{freq_{c_i,t}\times freq_{total}}{freq_{c_i}\times freq_t} \tag{3}$$

where t denotes the target word and c_i denotes the associated word neighboring t in the context window, and in the experiments of this chapter we preset the length of the context window to 5. $freq_{c_i,t}$ denotes the number of times the associated word c_i co-occurs with the target word t, $freq_{total}$ denotes the number of occurrences of all words, $freq_{c_i}$ denotes the number of occurrences of contextual words, and $freq_t$ denotes the number of occurrences of the target word.

In the experiment, in order to avoid the interference of negative information on the real semantic information, we choose positive point-by-point mutual information (PPMI) to calculate the semantic information of words, and its specific calculation process is shown in Equation (4):

$$ppmi(c_i,t) = \begin{cases} pmi(c_i,t) & \text{if } pmi(c_i,t)\geq 0 \\ 0 & \text{else} \end{cases} \tag{4}$$

II. B. 2) Word2vec to get word embeddings

Word2vec is a commonly used method for obtaining embedded representations of words [15]. Both models realized by Word2vec contain three layers, namely, input layer, mapping layer, and output layer, but the direction of their input and output is just opposite.

Among them, CBOW model mainly predicts the target word through contextual word information [16]. In CBOW model, all the words in the context have equal weights on the probability of occurrence of the current word, so the order of occurrence of the words in the context is not taken into account. Skip-gram model mainly predicts the distribution of words in the context through the current word. In the Skip-gram model, Skip ensures that words within a predefined window will have their probabilities calculated even if there are some words in between, thus effectively solving the interference of structural auxiliaries and so on.

In addition, Word2vec also provides two sets of optimization algorithms to improve the training efficiency of word vectors, namely, Hierarchical Softmax and Negative Sampling, in which Hierarchical Softmax uses Hoffman coding according to the word frequency, and Hoffman coding can be utilized to carry out the hierarchical word frequency features. Negative Sampling does not refresh the weights of all nodes, but selects some nodes labeled as 0 to refresh the weights, which effectively reduces the complexity of the algorithm.

The input layer of the Skip-gram model is the word vector corresponding to the current word, and the constant mapping is carried out in the mapping layer [17]. However, finally, in the output layer, the word frequency in the corpus is used as the weight to construct a Hoffman tree, and the mapping layer is connected to the non-leaf nodes

of the Hoffman tree, the leaf nodes correspond to all the words in the vocabulary list, and any one of the non-leaf nodes are vectors, but they do not represent a certain word but a certain class of words, assuming that the leaf nodes and the non-leaf nodes all correspond to the same vector, in this case, we consider that the vector corresponding to the leaf nodes is the word vector. In this case, we consider the vector corresponding to the leaf nodes as the word vector, and the vector corresponding to the non-leaf nodes as the auxiliary vector. Therefore, the vectors corresponding to all leaf nodes are the word vectors that we have trained.

II. C.Context-dependent phrase-pair semantic matching models

The context-dependent phrase-pair semantic matching model proposed in this paper aims to dynamically calculate the translation probability between source language phrases and different target candidates based on the context. The basic framework of the context-dependent phrase pair semantic matching model is shown in Figure 1. In this section, the method of calculating the semantic matching degree of phrase pairs in a given context is described according to this figure. Specifically, it is divided into the following three main parts: first, it introduces the sentence semantic feature extraction method based on convolutional neural network; second, it introduces the fusion strategy of sentence semantics and document topic information; third, it introduces the computation method of the semantic matching degree between source language phrases and target phrases in the context.

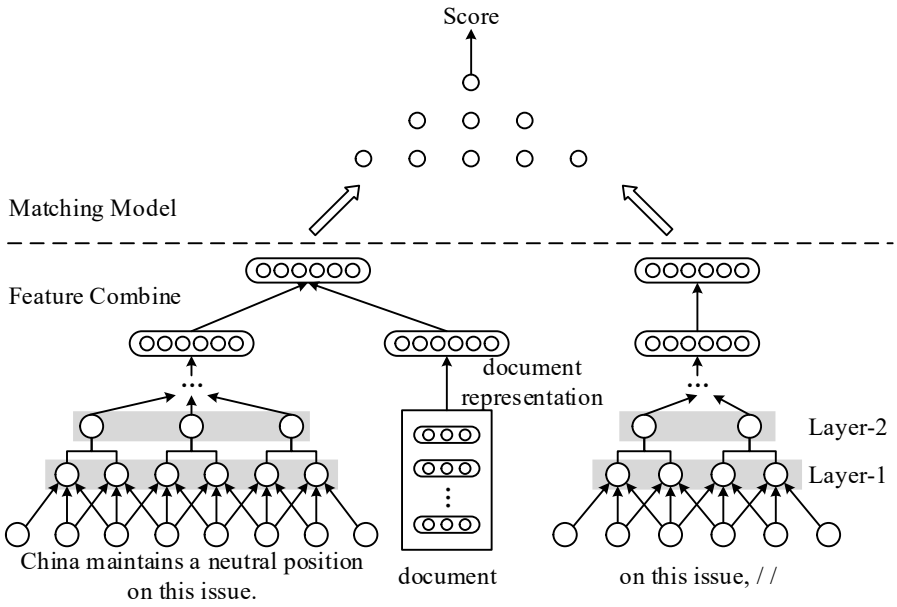

Figure 1: Context-dependent phrase-pair semantic matching model framework

II. C. 1) Convolutional Neural Network-based Sentence Feature Extraction

In this paper, we use convolutional neural networks to learn semantic representations of sentences and candidate target phrases. For the input sentence, the words in the sentence are first initialized into pre-trained word vectors to obtain the word vector matrix of all the words in the sentence. Secondly, the convolutional neural network performs convolutional operations on all possible windows of the input sentence and pools the features obtained from the convolution for selection. After several convolution and pooling operations, the semantic feature representation of the sentence is finally obtained. Assuming that the length of the input sentence is n and $x_i \in R^k$ is the word vector of the i th word in the sentence (with dimension k), the source language sentence can be viewed as a matrix of dimension $n \cdot k$. For ease of presentation, the sentence is represented in the form shown in equation (5):

$$x_{1:n} = x_1 \oplus x_2 \oplus \cdots \oplus x_n \tag{5}$$

where $\oplus$ denotes the concatenation operator and $x_{i:i+j}$ denotes the concatenation of word vectors of the words $x_i, x_{i+1}, \cdots x_{i+j}$.

In the first layer (Layer-1), the convolutional layer of the neural network takes as input a matrix of word vectors for a source language sentence f or a candidate phrase $\hat{e}$, and operates a sliding window of length h to perform a convolutional transform on all possible combinations of neighboring word sequences using a filter

$w \in R^{h \cdot k}$. For each possible combination of word sequences, a new feature representation is generated, which is eventually combined into a feature vector for the sentence or phrase. This is shown in equation (6):

$$c_i^{(1,j)} = f\left(w^{(1,j)} \cdot x_{i:i+h-1} + b^{(1,j)}\right) \tag{6}$$

where $c_i^{(1,j)}$ denotes the features generated by the convolution of the window $x_{i:i+h-1}$ through the j th filter, $w^{(1,j)}$ denotes the parameter matrix corresponding to the j th filter in the first layer, $b^{(1,j)} \in R$ denotes the bias term, and f denotes the nonlinear activation function, and in this paper, the ReLU function is used as the activation function.

In order to distinguish phrases and their contexts, this paper adds a new dimension after the word vector, which is used to distinguish whether the current word is within the scope of the phrase pair or not, where 1 means that the word is inside the phrase pair and 0 means other context words.

In the second layer (Layer-2), its input is the feature vector obtained from the convolution of the previous layer. In this paper, all adjacent non-overlapping convolutional features are pooled for selection. The details are shown in Equation (7):

$$c_i^{(2,j)} = max\left\{c_{2i}^{(1,j)}, c_{2i+1}^{(1,j)}\right\} \tag{7}$$

After several convolution and pooling operations, this paper obtains the feature vectors of source language sentences and target phrases.

II. C. 2) Strategies for integrating sentence semantics and document topics

This section first introduces the document topic feature acquisition method, and then, introduces the fusion strategy of sentence semantic features and document topic features.

(1) Document topic feature representation

In this paper, we use the Wikipedia document set to train the LDA topic model and use the generated model to reason about the topic vectors of all documents in the translation system training set. The topic vectors of the documents represent the probability that a document belongs to each topic. This topic information will be used to aid translation decisions.

Suppose a document is represented as an ordered combination of all word lexical vectors, i.e., $d = \{d_1, d_2 \cdots d_n\}$. Then the vector representation of the document is shown in Equation (8):

$$V = \frac{\sum_{i=1}^{n} w(i) d_i}{\sum_{i=1}^{n} w(i)} \tag{8}$$

where V denotes the semantic vector of the document, $w(\cdot)$ denotes the weight of the i th word in the document, in this paper, we use the TF-IDF metrics as the weighting function, and d_i denotes the word vector of the i th word in the document.

(2) Different fusion strategies

In this paper, we propose to utilize both simple splicing and shallow neural networks to fuse the information of sentences and documents, so as to make full use of the information of both and calculate the semantic match between source language phrases and target candidate phrases in context.

II. C. 3) Phrase Pair Semantic Matching Calculation

Based on the above method, this paper obtains the fused semantic representations of source language phrases and context information as well as the semantic representations of target candidate phrases. On this basis, this paper uses a multilayer perceptron to calculate their semantic matching degree, which indicates the probability of a source language phrase being translated into a target candidate phrase in a given context.

In this paper, the semantic feature vectors of the source and target languages are spliced together as inputs to the multilayer perceptron. The first layer of the multilayer perceptron machine first nonlinearly varies the spliced feature vectors to get the hidden layer state. The details are shown in Equation (9):

$$h_c = \phi(w_c \cdot [s_{\tilde{f}_i} : t_{\tilde{e}_j}] + b_c) \tag{9}$$

where $s_{\tilde{f}_i}$ denotes the fused semantic features of the source language phrase in context and $t_{\tilde{e}_j}$ denotes the semantic features of the target candidate phrase.

The second layer of the multilayer perceptron takes the above hidden layer as input, gets a new hidden layer after a nonlinear transformation, and reaches the output layer after another linear transformation. The output layer has only one node, which represents the semantic matching score of the phrase pair in the context. The specific formula is shown in below:

$$score(s,t) = W_{f_2}\left[\phi(w_{f_1}\cdot h_c + b_{f_1})\right] + b_{f_2} \tag{10}$$

II. C. 4) Training Strategies for Semantic Matching Models

The training goal of the context-dependent phrase-pair semantic matching model proposed in this paper is to assign high matching scores to correct translation results of source language phrases in specified sentence and document contexts, and low matching scores to incorrect translation results. Based on the idea of pairwise ranking learning, this paper transforms the problem of ranking all candidate translations of a source language phrase into a problem of ranking two-two candidate translation pairs.

The pairwise ranking loss function is adopted as the loss function of the model, as shown in Eq. (11):

$$L_\theta = max\left\{0, 1 + score(s,t^-) - score(s,t^+)\right\} \tag{11}$$

where $score(s,t)$ denotes the semantic matching degree of a phrase pair defined according to Eq. (11). θ denotes all the parameters of the neural network model, including the sentence and target phrase convolutional network (two-layer convolution, two-layer pooling), the shallow network for fusion of sentence and document information, and the parameters of the multilayer perceptron.

In this paper, we use the above methods to construct the training dataset and minimize the target loss function training to obtain the context-dependent phrase pair semantic matching model. The loss function on the full training set is defined as shown in Equation (12):

$$L = \sum_{(s,t^+,t^-)\in C} L_\theta(s,t^+,t^-) \tag{12}$$

In this paper, we use mini-batch gradient descent algorithm to optimize the parameters of the neural network model. When translating and decoding, the trained model and the context in which the phrase is located are used to determine the semantic match between the current phrase and all target candidate phrases, and are incorporated into the machine translation system as new features.

III. Translation model performance testing

In this paper, we use a processed Chinese parallel corpus of 600,000,000, which is randomly divided into 580,000,000, 10,000,000 and 10,000,000 as the training set, validation set and test set respectively, to test the English translation performance of the proposed translation model based on the semantic distribution of words in this paper.

In the four experiments conducted with random seeds of S_1=64, S_2=512, S_3=2048, S_4=6132, respectively, the BLEU performance of this paper's translation model is compared with those of the comparative Transformer_Base model, Transformer _Syn model as a comparison of the BLEU value comparison results are shown in Table 1. It can be seen that this paper's model outperforms both Transformer_Base model and Transformer_Syn model in terms of BLEU values, with an improvement of 2.06 and 5.12, respectively.

Table 1: Comparison of experimental results

Model	Random seed				Mean value
	S_1	S_2	S_3	S_4	
Transformer_Base	38.92	38.75	39.01	38.89	38.9
Transformer_Syn	41.42	42.45	40.97	43.01	41.96
Model in this paper	44.51	43.78	45.41	42.38	44.02

The relationship between the training time and the number of parameters for each model is specifically shown in Table 2. Although the model in this paper is slightly higher than the other two models in terms of the total number of parameters, with an increase of 2401755 and 3290311, respectively, in combination with the above its significant

improvement in BLEU values illustrates the effectiveness of the additional parameters. These additional parameters provide the model with more learning capacity, enabling it to better understand and process complex syntactic structures. In terms of training time, although it takes longer, this extra computational cost is justified given the significant improvement in model performance. To summarize, the translation model based on word semantic distribution proposed in this paper achieves a significant performance improvement in translation tasks.

Table 2: Model parameters and training time

Model	Total parameters	S_1	S_2	S_3	S_4	Average time (min)
Transformer_Base	50131065	1624	1598	1635	1644	1625.30
Transformer_Syn	51019621	2125	2195	2272	2098	2172.60
Model in this paper	53421376	2396	2513	2465	2597	2492.80

In this paper, we take the random seed S_1=64 and compare it with other translation model related works such as PASCAL, Localness, Context-Aware, Lisa, etc., and the experimental results are shown in Table 3. As can be seen from the table, the BLEU value of the model in this paper is the highest, 44.51, and its training time is 2396 minutes, which has a significant time advantage over the Context-Aware model (2875 minutes), although it is slightly longer than the Localness model (1986 minutes), which indicates that the model proposed in this paper maintains a high BLEU score while also maintaining a certain level of training efficiency. The model in this paper outperforms all other compared models in terms of translation quality. Although its training time of 2,396 minutes is not the shortest, this training time can be considered as a reasonable compromise between efficiency and performance considering the performance improvement.

Table 3: Comparison of results from each model

Model	BLEU	Training time(min)
PASCAL	43.21	2113
Localness	42.56	1986
Context-Aware	42.48	2875
Lisa	41.72	3311
Model in this paper	44.51	2396

IV. Experiment on the English translation of the word “body” in the Analects of Confucius

In this chapter, we will use the translation model based on the semantic distribution of words proposed in this paper to translate the original text of the Analects into English, integrate the philosophical connotation of the word “body” in the Analects, and analyze the reader acceptability of the results of the English translation.

IV. A. Analysis of the Philosophical Connotation of the Character “Body” in the Analects of Confucius

The word “body” in the Analects is rich in philosophical connotations, encompassing not only concrete physical existence, but also extending to the dimensions of moral cultivation, social relations and the value of life. In the ideological perspective of the Analects, the body is not only the material material of human beings, but also the cohesive body of human nature, and the practice of benevolence and virtue is the process of humanization of the body and the carrying of the Way by the body. The philosophical connotation of the word “body” is summarized by combining the results of the English translation of the word “body” as translated by the translation model in this paper.

(1) Body-oriented: all moral practices begin with reflection on and cultivation of oneself.

(2) Dao-oriented: the individual's “body” ultimately needs to transcend the ego and be integrated into the Confucian order centered on “benevolence” and “propriety”.

Understanding the philosophical connotation of the word “body” in the Analects can help readers better understand the philosophical thoughts of the Analects and fundamentally grasp the philosophical qualities of the Analects.

IV. B. Readers' acceptability analysis of the results of the English translation of the word “body” in The Analects of Confucius

IV. B. 1) Objects of study

In this paper, the survey respondents-readers of the English translation of The Analects of Confucius are divided into Chinese readers and foreign readers. Satisfaction sampling from the readers of the translated text of The Analects, the most ideal readers among the selected survey respondents are those who know both the original language and the target language. In order to ensure the validity of the scoring, the readers' questionnaire for this study was distributed only among readers with a university degree or higher, who were required to be proficient in the target language and have a certain degree of bilingualism between English and Chinese. The Chinese readers came from teachers engaged in English-Chinese translation research in many Chinese universities, and most of them had overseas visiting scholar experience; the foreign readers came from foreign teachers and foreign students in China and students of Confucius Institutes in the United States, and their mother tongue was English. The details of the readers of the invited translations are shown in Table 4.

Table 4: Translation reader classification

Translated readers	Foreign readers	Chinese readers
Different points	Learning the original language	Proficient in primitives
	Academic research on non-English-Chinese translation	Academic research on English-Chinese translation
Common ground	With a bachelor 's degree or above (except for some students in Confucius Institute), proficient in the target language, and a certain ability to understand English and Chinese, it is an ideal target reader group.	

IV. B. 2) Research methodology

The questionnaire design criteria of this paper refer to the English-Chinese Machine Translation Test Syllabus of Peking University, and 10 translations (named V1~V10, respectively) are selected from the test corpus source obtained by applying the translation model based on the semantic distribution of words proposed in this paper. Manual scoring was used to allow the survey respondents to score and evaluate the translations at the semantic and main idea levels respectively.

In terms of survey practice and data processing, from designing the questionnaire to distributing the questionnaire to data processing, the experimental method was strictly followed. In this paper, the statistical tool SPSS was used to conduct statistics and analysis, and the scale data were analyzed for reliability and validity to ensure that the data were true, objective and valid.

IV. B. 3) Analysis of reader acceptability

(1) Semantic level

The acceptability of the semantic level means that the more the readers perceive the translation to be equal to the semantics of the original text in terms of form and meaning (the smaller the deviation), the higher the acceptability. On the contrary, the greater the semantic deviation between the translated text and the original text, the more errors there are, the lower the acceptability. In this paper, the semantic acceptability is set to five measures, with 1 being the lowest acceptability and 5 being the highest. The main idea acceptability of foreign readers and Chinese readers is specifically shown in Figure 2. The mean values of semantic acceptability of Chinese and foreign readers are 3.25 and 3.12 respectively, i.e., it shows that both of them show basic acceptance (2.5-3.5) of the semantic level, and the acceptability of Chinese readers is slightly higher than the semantic acceptability of foreign readers. The two curves scored by foreign readers and Chinese readers are basically the same, indicating that the semantic acceptability of the two has commonality and is meaningful for analysis. From the standard deviation of the two, the standard of foreign readers and the standard deviation of Chinese readers are almost on a straight line, and most of the standard deviation of the acceptability of the two is less than 1, which indicates that the two sets of data fluctuate less, and the data show a more obvious trend - basic acceptance of semantics.

(2) Main idea level

The acceptability of the main idea level refers to the reader's acceptance of the main idea of the translated text, i.e., the more the translated text is equal to the main idea of the original discourse in form and meaning (the smaller the deviation), the higher the acceptability. On the contrary, the more the translation deviates from the main idea of the original discourse and the more errors there are, the lower the acceptability. Specifically, the scoring mainly focuses the readers on the fidelity of the main idea, and scores and evaluates whether the main idea of the translation of the Analects is equivalent to the main idea of the original text in terms of the translation of the subject matter of the discourse and the translation of the terms, respectively.

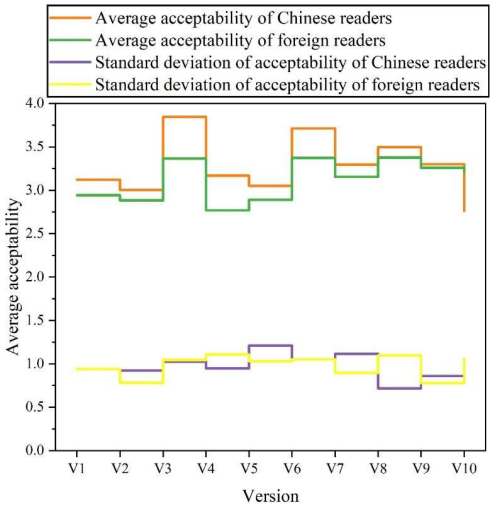

Figure 2: Semantic acceptability of foreign readers and Chinese readers

In this paper, we also set the degree of acceptability of the main idea into five measures, with 1 being the lowest acceptable degree and 5 being the highest acceptable degree. The main idea acceptability of Chinese and foreign readers is shown in Figure 3. The mean values of Chinese and foreign readers' main idea acceptability are 3.32 and 3.15 respectively, i.e., it shows that both of them show basic acceptability (2.5-3.5) to the main idea level, and the acceptability of Chinese readers is slightly higher than that of foreign readers. From the standard deviation of the two, the standard of foreign readers and the standard deviation of Chinese readers are almost on a straight line, and most of the standard deviation of the acceptability of the two is less than 1, which indicates that the two sets of data fluctuate less, and the data show a more obvious trend - basic acceptance of the main idea.

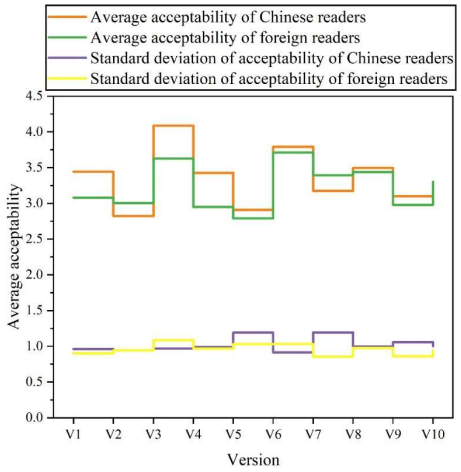

Figure 3: The acceptability of the gist of foreign readers and Chinese readers

(3) Overall acceptability of the translation

MT acceptability, i.e., the overall acceptability of the machine translation, refers to the readers' overall acceptance of the machine English translation of the Analects, which is the readers' evaluation of the overall acceptability given by the translation.In this study, in order to test the hypotheses, we set up a measure of the overall acceptability, i.e., the MT acceptability (measured value).The MT*acceptability (computed value) is the average of the word meanings and the main idea acceptability of words. This is shown in Table 5. As can be seen from the table, in terms of the standard deviation of the two, the MT acceptability and MT* acceptability of foreign readers are smaller than those of Chinese readers, indicating that the data volatility of the overall acceptability of foreign readers is smaller than that of Chinese readers. The overall acceptability and the acceptability scores of the semantic and main idea levels of both are above 3, and the acceptability level is basically acceptable. However, the acceptability of Chinese

readers is higher than that of foreign readers in both semantics and main idea, indicating that Chinese readers are more acceptable than foreign readers in both machine translation. In addition, the calculated values of MT* acceptability for foreign readers and Chinese readers are very close to the measured values of MT acceptability, and the absolute value of the error is less than 0.15.

Table 5: Overall acceptability of the translation

Acceptability	Foreign readers		Chinese readers	
	Average value	Standard deviation	Average value	Standard deviation
Semantic acceptability	3.12	0.56	3.25	0.66
The acceptability of the subject matter	3.15	0.48	3.32	0.75
MT Acceptability (measured)	3	0.57	3.3	0.76
MT * Acceptability (calculated)	3.14	0.52	3.29	0.67

Overall, the translation model based on the semantic distribution of words proposed in this paper is able to translate the Analects of Confucius better, and the acceptability of the resulting translations of the Analects of Confucius by foreign and Chinese readers is basically acceptable.

V. Conclusion

This paper proposes a translation model based on the semantic distribution of words to provide technical support for the English translation of the Analects. The translation performance of this paper's translation model is tested. In four experiments with random seeds of S_1=64, S_2=512, S_3=2048, S_4=6132, this paper's model improves the average BLEU value by 2.06, 5.06, 5.0, 5.0 and 5.0 respectively, compared to the comparative Transformer_Base model and Transformer_Syn models by 2.06 and 5.12, respectively.The total number of parameters and training time are slightly higher than the other models, but this extra computational cost is reasonable given the significant improvement in model performance. The random seed S_1=64 is selected for comparison experiments with other translation models such as PASCAL, Localness, Context-Aware, Lisa, etc. The model in this paper has the highest BLEU value of 44.51, and the training time is 2,388 minutes, comparing with the Context-Aware model (2,875 minutes) There is an obvious time advantage, slightly longer than the Localness model (1986 minutes). The model in this paper is far superior to the other comparative models in terms of translation quality. Although the training time is slightly longer, based on the performance improvement of the model itself, the training time can be regarded as a reasonable compromise between efficiency and performance.

Using the translation model based on the semantic distribution of words proposed in this paper to translate the original text of The Analects of Confucius into English, an English translation experiment is conducted to explore the reader acceptability of the resulting English translation results. At the semantic level, the mean values of semantic acceptability for Chinese and foreign readers are 3.25 and 3.12 respectively, both of which show basic acceptance (2.5-3.5). At the main idea level, the mean values of main idea acceptability for Chinese and foreign readers are 3.32 and 3.15 respectively, also showing basic acceptance (2.5-3.5). As far as the overall acceptability of the translation is concerned, the MT acceptability (3.14) and MT* acceptability (3.29) of foreign readers are smaller than those of Chinese readers. At the same time, the calculated values of MT* acceptability and the measured values of MT acceptability for foreign readers and Chinese readers are very close to each other, and the absolute value of the error is less than 0.15.On the whole, the translation model based on the semantic distribution of words proposed in this paper is able to realize high-quality English translation of The Analects of Confucius, and the results of the English translation can be accepted by both Chinese readers and foreign readers.

References

[1] Peimin, N. (2024). The Analects as a Philosophical Work. In The Civilization of China and the Civilizations of the World (pp. 149-171). Singapore: Springer Nature Singapore.

[2] Min, F. (2021). A critical study on translation of the analects: An ideological perspective. International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL), 3(1), 45-54.

[3] Connolly, T. (2025). The Aspiring Confucian: Long-Term Transformation in the Analects. Journal of Chinese Philosophy, 51(2-3), 113-122.

[4] He, M. (2017). A comparative multidimensional study of the English translation of Lunyu (The Analects): A corpus-based analysis. GEMA Online Journal of Language Studies, 17(3), 37-54.

[5] Hou, Y., & Sun, Y. (2019). A Corpus-Based Comparative Analysis of Cohesive Devices in Two English Translations of The Analects of Confucius. International Journal of Languages. Literature and Linguistics, 5(4), 247-252.

[6] Yang, D. (2022). An Investigation on English Translations of Culture-Loaded Words in The Analects of Confucius from the Eco Perspective: A Case Study of the English Translation of Lectures on China’s Traditional Political Thoughts. Editorial Board, 7.

[7] Shi, Y., & Hu, W. (2024). A Study on the Translation of Culturally Unique Chinese Terms Based on Corpora—Taking English Translations of The Analects as an Example. Journal of Linguistics and Communication Studies, 3(4), 43-47.

[8] Yang, L., & Zhou, G. (2024). Dissecting The Analects: an NLP-based exploration of semantic similarities and differences across English translations. Humanities and Social Sciences Communications, 11(1), 1-12.
[9] Bei, L. (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, 2020(1), 8827657.
[10] Wang, Q. (2022). Semantic analysis technology of English translation based on deep neural network. Computational Intelligence and Neuroscience, 2022(1), 1176943.
[11] Bi, S. (2020). Intelligent system for English translation using automated knowledge base. Journal of Intelligent & Fuzzy Systems, 39(4), 5057-5066.
[12] Camacho-Collados, J., Pilehvar, M. T., Collier, N., & Navigli, R. (2017, August). Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017) (pp. 15-26).
[13] Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. Nature Human Behaviour, 4(10), 1029-1038.
[14] Jing Li. (2024). English-Chinese Translation Quality Assessment Based on Phrase Statistical Machine Translation Decoding Algorithm. International Journal of Maritime Engineering,1(1),675-688.
[15] Zhan Zerui. (2025). Comparative Analysis of TF-IDF and Word2Vec in Sentiment Analysis: A Case of Food Reviews. ITM Web of Conferences,70.
[16] Bing Liu. (2020). Text sentiment analysis based on CBOW model and deep learning in big data environment. Journal of Ambient Intelligence and Humanized Computing,11(2),451-458.
[17] Enes Celik & Sevinc Ilhan Omurca. (2024). Skip-Gram and Transformer Model for Session-Based Recommendation. Applied Sciences,14(14),6353-6353.