Design of Networked Intelligent Translation System Based on Machine Learning Algorithm

Liu Hong*
Wuhan Donghu University, Wuhan, 430070, China

*Corresponding author: fanjxlve896@163.com

Abstract. China's social and economic development is relatively rapid, and China's opening to the outside world is further expanding. Under this background, the demand for foreign language professionals is also expanding. This paper designs a networked intelligent translation system based on machine learning algorithm, and proposes a cross-language information extraction method based on bilingual word representation. The learning of bilingual word representation can be divided into two stages: unsupervised and supervised. Capturing Chinese and English bilingual semantic information. The cross-language information extraction method based on machine learning algorithm reduces the impact of translation errors and language gaps on cross-language classification performance, obtains effective bilingual semantic information and annotation information, and improves the cross-language information extraction performance. The test results show that the designed automatic English translation system can effectively and quickly realize intelligent English translation of memory-assisted long characters, with high data recall rate, good accuracy and reliability, and good system compatibility, which is worthy of popularization and practical application.

Keywords: Machine learning; Network intelligent translation system; Cross-language information extraction

1. Introduction
Translation is the only way to transform multiple languages, and through translation, we can communicate clearly what both sides express. For Chinese and English, language communication adopts basic translation to realize mutual conversion, and numbering conversion is carried out according to slogan grammar and semantics. The main advantages of electronic dictionary are intelligence, comprehensive functions and portability. The mobile electronic dictionary in Android system enables Chinese people to learn foreign languages better [1]. With the continuous development of mobile networks, mobile terminals have gradually replaced Internet terminals, which have developed rapidly in modern society, so their potential can not be estimated, and the development space of mobile terminal application software is relatively large. The design of networked intelligent translation system is based on the design of machine translation algorithm [2-3], which combines artificial intelligence
algorithm to design networked intelligent translation software, carries out cross-compilation control and program loading of automatic translation in embedded environment, realizes the combination test of phrases and translations, and improves the intelligent translation ability.

The translation of existing machine translation system is difficult to guarantee complete accuracy, and it usually needs manual intervention and modification to form an accurate translation. Therefore, we propose a design scheme of networked intelligent translation system based on machine learning algorithm [4]. This platform realizes integrated processing of multiple translation modes based on rule analysis, analogical reasoning and statistical knowledge, provides man-machine interaction interface, realizes manual intervention of translation results, realizes object-oriented engineering [5], task and user management, and realizes organic integration of various translation strategies, thus improving the translation quality and speed of practical translation systems.

2. Algorithm design of intelligent translation system

2.1 Algorithm design of intelligent translation system

Bilingual word representation lacks contextual information of words, and the unsupervised learning stage of bilingual semantic information learning is completely independent from the supervised learning stage of tagging information learning. A cross-language information extraction method based on joint representation learning is proposed. Long-term and short-term memory recursive network (LSTM) is used to learn Chinese and English bilingual representation. On the basis of semantic representation of words, contextual emotional (fuzzy) information representation is introduced, and semantic and emotional (fuzzy) information of emotional words (hedges) in specific contexts are jointly trained [6]. At the same time, bilingual semantic learning and tagging information learning are combined into a learning stage to further optimize the performance of cross-language information extraction.

In the task of emotional classification, the context information of emotional words plays an important role in judging the emotional tendency of documents. In this paper, CSIR is obtained based on LSTM training (Figure 1).

![Fig.1 Schematic diagram of obtaining SSR based on LSTM training](image)

Firstly, taking the emotional word $senti_i$ as the center and the window as [-2, 2], the context sequence of Chinese and English emotional words is extracted, and the emotional fragment $x = \{word_{i-2}, word_{i-1}, senti_i, word_{i+1}, word_{i+2}\}$ is obtained.

Then, the 50-dimensional word representation of the obtained sequence is taken as the input of LSTM at each time $t$, and the hidden representation $h_t$ at $t$ time is calculated and obtained.

Then, the hidden representations at all times are averaged and pooled to obtain 50-dimensional hidden representations $h$ of emotional fragments, which we call segment sentiment representations.
(SSR). Input $h$ into the logistic regression layer, and calculate the classification loss based on the emotional polarity of emotional words.

Finally, the model parameters are updated based on Adadelta algorithm [7], and SSR is obtained by training.

In this paper, three classification losses are defined based on log-likelihood function: $l_E$, $l_C$ and $l_{E-C}$. $l_E$ is hidden in English for classification loss of $h_E$, $l_C$ is hidden in Chinese for classification loss of $h_C$, and $l_{E-C}$ is hidden in Chinese and English for classification loss of connecting $\{h_E, h_C\}$. Add the above three classification losses as the classification loss $L_{\text{pred}}$ of the model, as shown in formula (1).

$\text{CE}_{\text{pred}} = l_E + l_C + l_{E-C} \quad (1)$

In addition, according to the related research of neural network machine translation, LSTM can encode the vector of the source language into hidden representation and decode it into the vector of the target language. We think that the difference between hidden representation $h_E$ and $h_C$ should be as small as possible, so as to reduce the vector difference between Chinese and English. Therefore, bilingual semantic loss $L_{\text{sem}}$ is defined. Reduce the gap between the two languages, as shown in formula (2).

$L_{\text{sem}} = \|h_E - h_C\|^2 \quad (2)$

In this paper, the weighted sum of classification loss and bilingual semantic loss is calculated as the objective function of the whole model, as shown in formula (3), where $\alpha$ is the weight of classification loss, and the model parameters are learned based on Adadelta algorithm.

$L = \alpha \cdot L_{\text{pred}} + (1 - \alpha) \cdot L_{\text{sem}} \quad (3)$

2.2 Information gain

Using the method of information gain to select effective style features can reduce the complexity of the experiment, save the experimental time and improve the accuracy [8]. In this selection method, the criterion to measure the importance of features is the amount of information that features can provide for the classification system. The more information provided, the more important the features are, and the information can be calculated by information entropy. Information entropy is defined as:

$H(X) = -\sum_{i=1}^{n} P(X_i) \log P(X_i) \quad (4)$

For the classification system, category $C$ is a variable. If $C$ has $n$ features, the possible values of $C$ include $C_1, C_2, \cdots, C_n$; Assuming that the probability of each category is $P(C_1), P(C_2), \cdots, P(C_n)$, the entropy of the classification system can be expressed as:

$H(C) = -\sum_{i=1}^{n} P(C_i) \log P(C_i) \quad (5)$

Information gain is specific to specific information characteristics. For example, for a certain feature $t$, the amount of information when the system has or does not have the feature $t$ is calculated separately, and the difference between the two is the amount of information provided by the feature for the system, that is, the gain.

When the system has the feature $t$, the calculation of its information amount is shown in formula (5), which indicates the information amount when the system has all the features. If there is no feature $t$ in
the classification system, then the conditional entropy at this time is:

\[ H(C|T) = \sum_{c \in C} P(c) \times H(c|T) = \sum_{c \in C} P(c|T) \times P(T) \times P(c) \times P(T|c) \]

(6)

The difference between the two is the information gain of feature \( t \):

\[ IG(t) = H(C) - H(C|T) \]

(7)

According to the order of information gain from high to low, we extracted 9 effective classification features.

2.3 Optimization of translation algorithm

The important role of the concrete design of the system operation flow is to ensure the stability and reliability of the system operation. Therefore, the system starts the online translation mechanism first, and specifies the number of software operations. When running for the first time, the thesaurus is directly loaded and stored in a given folder. Instead of running for the first time, enter the main running interface and display the main functions such as word translation, word book, word query and word management.

Based on the binary semantic feature relevance extraction method, the English vocabulary of memory-assisted long characters is analyzed in detail, and the text is extracted from it to obtain the text similarity between memory-assisted long characters \( X \) and \( Y \), namely:

\[ \sin(X, Y) = \cos(X, Y) = \frac{T(X) \cdot T(Y)}{|T(X)| \cdot |T(Y)|} \]

(8)

Using the position of long-character English words in the text as the carrier, the context is adaptively matched to obtain the fuzzy concept set. According to the contextual thinking and attribute fields embodied in long-character English vocabulary, the concept set of translation output vector is modified moderately, so as to obtain the relevance between the semantic word length and the part of speech of the text [9]. According to the specific position of the text, the context self-service matching is realized, so as to obtain the feature quantity of inter-vocabulary interactive information. According to the feature feedback of interactive information between words, the translation is adjusted appropriately, the translation algorithm is optimized, and the translation rule calculation result is obtained, namely:

\[ \cos(X, Y) = \frac{T(X) \cdot T(Y)}{|T(X)| \cdot |T(Y)|} \]

(9)

Based on translation rule function, automatic English translation of memory-assisted long characters is realized.

3. Overall structural design

The architecture design of networked intelligent translation system based on machine learning algorithm is shown in Figure 2.

The system consists of user application module, translation module and system management module. The user application module provides login and translation services for users, and the translation module realizes the translation of words/sentences among various natural languages and presents feedback results to users. The system administrator queries and modifies translation rules through the management module.

Translation module is the core of networked intelligent translation system based on machine learning algorithm, and its architecture diagram is shown in Figure 3.

When the user uses the system, the translation module transmits the translation request to the server in the form of word vector after inputting the translation request on the translation interface. The server
uses the neural network translation model to translate words/sentences, which can set the limit of access times and complete the allocation of concurrent requests.

![Architecture diagram of machine automatic translation system](image1)

Fig.2 Architecture diagram of machine automatic translation system

![Architecture diagram of translation module](image2)

Fig.3 Architecture diagram of translation module

4. System function module

4.1 Image segmentation
Correct segmentation can get the correct recognition object, so as to effectively realize image analysis. If correct segmentation is not realized, then the following work is meaningless. Therefore, image segmentation has great significance in the system. In order to get single characters and words effectively, the general method is to realize horizontal projection and vertical projection of document images. Vertical projection can show the gap between characters, realize gap threshold setting and realize multi-character segmentation.

In the process of cutting overlapping characters, connected domain method and winding method can be used, but this algorithm is complex, time-consuming and has a large amount of calculation, so it can not be used in this system. Therefore, overlapping characters can be cut by reusing broken characters. Among them, the shaded part belongs to the overlapping part of characters, which will not lead to accumulated errors, so it should be fully considered when recognizing, thus reducing the recognition accuracy. In order to put forward this shadow part, it can be extended to make it a slightly broken character.

4.2 Translation information and its representation of source translation
The source translation information is used to express the source translation and its translation relationship information, in which the source text information is represented by a structure array of source text meaning segments, each meaning segment is an element in the array, and each array element is composed of a triple. The first member of the triple is the most important distinguishing feature of the source text meaning segments.

In general, the feature is expressed in the original text string, but it may also be an abstraction of the string at a certain level. The second member represents the feature attribute of the meaning paragraph,
and the third member represents the subscript (sequence) corresponding to the corresponding translation component. Translation information is also represented by a structure array of translation meaning segments expressed by a triple sequence, in which the first two attributes are the same as the source information expression, and the third member is the array subscript of the source meaning segment corresponding to the meaning segment.

The sequence of subscripts shows the correspondence of the source text, the omitted components in the translation and the new components in the translation. Moreover, the multiple mapping of subscripts in the source text can better find the indirect correlation between the components of the source text (or the components of the translation), which is helpful to accurately describe the complex translation correspondence and other knowledge.

4.3 Implementation of business logic layer

(1) Implementation of business logic layer

There is a typical hierarchical architecture: MVC(Model View Controller) architecture. Among them, model represents data, view represents interface, and controller represents business logic. This is similar to the hierarchical structure adopted in this system in principle. MVC architecture emphasizes that the change of business logic causes the change of data, and the change of data causes the change of interface. That is to say, business logic controls the behavior of the interface. This is contrary to the normal idea that the upper layer controls the lower layer behavior. Usually, in the interface of data-driven functions, this is easy to happen.

Data needs to be transferred between business logic and interface. If it is simple data, there is no problem. If it is a class object containing multiple types of data, if the entity class of business logic is directly used in the interface, it will cause coupling between the interface and business logic. The function trigger points in the interface will call the API in the business logic, pass parameters to the API, and obtain the execution result of the API, which is also a kind of coupling. But there is no way to avoid this coupling. The way to solve this kind of coupling is that the interface layer defines its own data structure independently, and an intermediate layer is added to take charge of the mapping between interface data structure and entity class. In this way, the entity objects in the data access layer will not be involved in the interface layer.

(2) Communication between business logic layer and other layers

The communication mode between business logic layer, data access layer and interface layer is shown in Figure 4.

Fig.4 Communication mode between business logic layer and data access layer and interface layer

The dynamic link library generated by the business logic layer and the dynamic link library generated by the data access layer will be deployed on the same physical node, so they can communicate with each other through class methods and method calls. There is one-way communication between business logic layer and data access layer, that is, business logic layer uses classes and methods in data access layer, but data access layer does not depend on business logic layer.

TCP communication is adopted between the business logic layer and the interface layer for message transmission. The transfer process is shown in Figure 5.
Message transfer process between business logic layer and interface lay

Figure 5 describes the one-way transmission process of data from interface to business logic, in order to simplify the message transmission process. From business logic to interface, data can be processed in reverse process.

Serialized data will be sent to the destination as the payload of TCP message. At the destination, the payload of TCP message will be unpacked to obtain one-dimensional data. Then one-dimensional data is deserialized and filled into data objects, and finally the data objects are mapped to entity objects.

5. Test analysis
In order to demonstrate the overall performance of networked intelligent translation system based on machine learning algorithm, further experimental verification analysis is carried out. With WINCC6.0 as the carrier, an experimental platform of memory-assisted English automatic translation system for long characters is constructed, and the performance of the system is tested comprehensively. The translation effects of the traditional system and this system are compared and analyzed, and the specific data recall rates of the two systems are clarified. The test results are shown in Figure 6.
It can be seen from fig. 6 that the data recall rate of the traditional system gradually increases with the increasing trend of English vocabulary, and the speed shows a trend of fast at first and then slow. When the vocabulary repository reaches 600Gibt, the data recall rate reaches the highest state, that is, 38%. However, with the increasing of English vocabulary, the data recall rate of this system gradually increases, and the speed is slow at first and then fast. When the vocabulary repository reaches 600Gibt, the data recall rate reaches the highest state, that is, 43%. It can be seen that the data recall rate of this system is obviously high, which means that the translation accuracy of the networked intelligent translation system based on machine learning algorithm is high, and it has good feasibility and practicality, which is worth popularizing and putting into application.

6. Conclusion
The main purpose of this paper is to establish a translation platform, integrate translation resources, improve the utilization rate of translation resources, establish an effective communication channel between translation talents and translation demanders, and provide high-quality and cost-effective translation services. This paper designs a networked intelligent translation system based on machine learning algorithm, and proposes a cross-language information extraction method based on bilingual word representation. The learning process is divided into unsupervised learning stage and supervised learning stage. In the unsupervised learning stage, the automatic coding machine for noise reduction is used for bilingual reconstruction in both Chinese and English to capture bilingual semantic information. The main advantage of this method is that it can save time and effort in feature extraction and language parameter discrimination calculation of massive languages, and can sort features according to discrimination. The results show that this system can quickly realize automatic translation of memory-assisted long characters, with good accuracy and reliability, and superior overall performance, which is worthy of practical application.

Acknowledgements
2020 Research Project of Humanities and Social Sciences of Department of Education of Hubei Province: Research on Training Mode of Innovative Talents of Translation Major Based on Critical Ability from the Perspective of Cognitive Taxonomy (Project Number: 20G068)

References
[1] Song X. Intelligent English translation system based on evolutionary multi-objective optimization algorithm. Journal of Intelligent and Fuzzy Systems, no. 10, pp. 1-11, 2020.
[2] Guo E. Application research of artificial intelligence English audio translation system based on fuzzy algorithm. Journal of Intelligent and Fuzzy Systems, no. 1, pp. 1-10, 2021.
[3] Bouarara H A, Hamou R M, Rahmani A. BHA2: Bio-inspired Algorithm and Automatic Summarisation for Detecting Different Types of Plagiarism. International Journal of Swarm Intelligence Research, vol. 8, no. 1, pp. 30-53, 2017.
[4] Li Wei, Chen Ruxu. Machine Learning Based Unstructured Large Data Analysis Algorithm for Communication Networks. digital technology and application, vol. 036, no. 010, pp. 136,138, 2018.
[5] Yuan Z, Jin C, Chen Z. Research on language analysis of English translation system based on fuzzy algorithm. Journal of Intelligent and Fuzzy Systems, vol. 40, no. 3, pp. 1-9, 2020.
[6] Li N. An improved machine learning algorithm for text-voice conversion. Journal of Intelligent and Fuzzy Systems, vol. 40, no. 2, pp. 1-11, 2020.
[7] Zhang Y. Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm. Journal of Intelligent and Fuzzy Systems, vol. 40, no. 2, pp. 2069-2081, 2021.
[8] Mara, Chinea-Rios, Germán, et al. Discriminative ridge regression algorithm for adaptation in statistical machine translation. Pattern Analysis & Applications, vol. 22, no. 4, pp. 1293–1305, 2019.
[9] Wei X. Simulation of English intelligent system based on CA-IAFSA algorithm and artificial intelligence. Journal of Intelligent and Fuzzy Systems, no. 2, pp. 1-11, 2021.