Research Article

Application of Feature Extraction Algorithm in the Construction of Interactive English Chinese Translation Mode

Guangjun Dong,1 Youchao Yang,2 and Qiankun Zhang3

1Anhui Vocational College of Electronics & Information Technology, Bengbu 233000, China
2Anhui University of Finance and Economics, Bengbu 233030, China
3Anhui Audit College, Hefei 230001, China

Correspondence should be addressed to Youchao Yang; 120081279@aufe.edu.cn

Received 19 August 2020; Revised 20 September 2020; Accepted 27 October 2020; Published 17 April 2021

Academic Editor: Hongju Cheng

Copyright © 2021 Guangjun Dong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the process of English translation, traditional interactive English translation system is not obvious in English semantic context. The optimal feature selection process does not achieve the optimal translation solution, and the translation accuracy is low. Based on this, this paper designs an interactive English Chinese translation system based on a feature extraction algorithm. By introducing the feature extraction algorithm, the optimal translation solution is selected, and the semantic mapping model is constructed to translate the best translation into English Chinese translation. The real experiment results show that the interactive English Chinese translation system based on feature extraction algorithm can get the best solution.

1. Introduction

In the process of daily English translation, an interactive English translation system is usually used for English Chinese translation [1, 2]. However, the traditional interactive English translation system does not have the best feature context selection in the process of translation feature semantics and context extraction, resulting in the low accuracy of translation [3].

In view of the above situation, this paper designs an interactive English Chinese translation system based on the feature extraction algorithm. The feature extraction algorithm is introduced to select the feature semantics, and the semantic ontology mapping model is established [4, 5]. The optimal solution is selected for the interactive English Chinese translation process, and the English Chinese translation process is finally realized through coding. In order to ensure the effectiveness of the interactive English Chinese translation system based on the feature extraction algorithm, a comparative simulation experiment is designed. The experimental data show that the interactive English-Chinese translation system based on the feature extraction algorithm in this paper can effectively carry out interactive translation between English and Chinese and solve the selection process of the optimal feature context [6, 7].

2. An Overview of Text Classification

With the development of electronic information technology, how to classify data accurately, efficiently, and quickly has become a hot topic at present. The feature word set is generated by multifeature extraction algorithm, the text feature vector is generated by the TF-IDF frequency algorithm, and the text classification is classified by the support vector machine (SVM) classifier model [8, 9]. And the corresponding calling interface is designed for the classification system, which ensures the usability of the classification module. At the same time, a Classified Thesaurus is designed to preserve the unique feature words of each category and to prioritize the categories of documents to be classified [10].

2.1. Two Methods of Text Representation. The vector space model of Salton is also called the word bag representation of text. The way to express text in bag words usually includes participle, taking root, removing function words, statistical
word frequency, selecting features, vectorization, and normalization [11, 12].

(1) Participle: according to certain rules, the text composed of strings is decomposed into a set of different words. By segmenting the training samples, a set of text representation Dictionary of all the different words in the training sample is obtained [13, 14]. (2) Take synonyms: there are many synonyms in Chinese documents, which express the same meaning and merge these synonyms into one word. (3) Functional words: pronouns, prepositions, auxiliary words, etc. are called functional words. Because functional words generally do not contain effective classification information, they can be eliminated from the dictionary of text representation. (4) Statistical word frequency: the frequency of every word root in a text representation dictionary in each text document. (5) Selection feature: select some words that contain significant category information in text representation dictionary. (6) Vectorization: vector models are used to represent each text in the form of vectors [15].

The classification system based on dictionary segmentation is influenced by the completeness of dictionaries and the segmentation algorithm. Aiming at the drawback of dictionary classification system, a text feature extraction technology based on N-gram information is proposed. It makes the text data classification system free from the dependence on the complex word segmentation processing program and the large word library and realizes the domain independence and time independence of the Chinese text data classification [16, 17]. The concept of N-gram information is a string proposed by the founder of information theory, Che-saning, in the study of source code, which is often used to represent the continuous n characters of the source output. Shannon has used it to study the statistical characteristics of characters or strings in English text, that is, the entropy of information. Then, N-gram information is widely used in the fields of text compression, character recognition, and error correction. It is a direct code-oriented technology [18, 19]. Classifier of Naive Baye is shown in Figure 1.

N-gram according to the byte stream to use the length of N of the window segmentation, such as People’s Republic of China, according to N = 4 sliding window segmentation called China, the people, Republic, country, such 5 grams. Because in Chinese, the two-word phrase has the greatest probability of occurrence, two words for four bytes, so N usually takes 4 [20].

In the VSM model, there are many ways to calculate the weight of entry TK. According to its value type, it can be divided into two kinds: the first Boolean type and its calculation method are to use all the entries of training documents as a complete set. If an entry TK appears in the document, its weight is 1; otherwise, it is set to 0. The second is the real number: the weight of the document is calculated according to the weight formula of the characteristic word [21, 22].

2.2. Commonly Used Classification Methods. The KNN method, one of the most famous pattern recognition statistics methods, has been used for more than 40 years and has been used in text classification research early. It is an example that is based on text classification method. For a test text, calculate the similarity between each text and its training set, and find the most similar K training texts based on text similarity. Then, each text class is scored based on the sum of documents similarity between the text and the test text in K training documents. Sort by value and specify the type of test text according to the score [23].

If we use all the sample points in the class as the representative points of the class, we call them the nearest neighbour method. It was proposed by Cover and Hart in 1968, and it is an important nonparametric method in pattern recognition [24]. The nearest neighbour method calculates the distance from the sample to all the representative points, and classifications are classified into categories belonging to the nearest representative point. In order to overcome the defects of the nearest neighbour method, the nearest neighbour is generalized to the k-nearest neighbour. It selects the nearest K representative point from the sample to be categorized to see which category of the K representative points belongs to the class and puts the test sample into the class. Linear separable schematic diagram is shown in Figure 2.

It can also be said that given a test document to be classified, the system finds the most similar k-nearest neighbours in the training set and gives the candidate category of the document based on the category of the nearest neighbour [25, 26]. The similarity between neighbour document and test document can be used as the class weight of the document in the nearest neighbour class. If some of the documents in the k-nearest neighbour belong to the same class, the sum of the class weights of each neighbour in the classification is the similarity between the category and the test document. By sorting the candidate class scores and giving a threshold, the categories of test documents can be determined [27].

Support vector machine, SVM, is a machine learning technology developed by Vapnik and its team led by Baer laboratories. The theoretical basis of SVM comes from the statistical learning theory put forward by Vapnik et al. The basic idea is that for a given learning task with a limited number of training samples, how to compromise the accuracy (for a given set of training sets), and the capacity of the machine (the machine can learn the ability of any training set without error) to get the best promotion performance. He adopted the principle of Structural Risk Minimization [28]. The SVM algorithm not only has a solid theoretical foundation but also achieves good results when applied to text categorization. If a vector of a training set can be linearly segmented by a hyperplane, and the distance between the nearest vector of the hyperplane is maximum, the hyperplane is called the best hyperplane. The closest point to the hyperplane is called
support vector. The point at which the decision surface is designed is called support vector (SV) [29].

Simple Bias classification is one of the commonly used methods in machine learning. The simple Bias classification method divides training instance \( I \) into feature vector \( W \) and decision category variable \( C \). The simple Bias classification assumes that the components of the eigenvectors are relatively independent relative to the decision variables, that is, each feature is independent of the category, which is the characteristic independence hypothesis [30]. For text categorization, it assumes that each word is independent between \( T_i \) and \( T_j \).

3. Feature Extraction

Feature extraction is a concept in computer vision and image processing. It refers to using a computer to extract image information and decide whether each image’s point belongs to an image feature. The result of feature extraction is to divide the points on the image into different subsets. These subsets often belong to isolated points, continuous curves, or continuous regions [31]. A method of transforming a set of measured values of a pattern to highlight the representation characteristics of the pattern. Image analysis and transformation are used to extract the required features [32]. Feature extraction refers to the method and process of extracting characteristic information from images by computer.

3.1. Basic Concepts. There is no universal and precise definition of the features so far. The precise definition of characteristics is often determined by questions or application types. Feature is an interesting part of a digital image. It is the starting point of many computer image analysis algorithms. Therefore, the success of an algorithm is often determined by the characteristics that it uses and defines. Therefore, one of the most important features of feature extraction is “repeatability:” the features extracted from different images of the same scene should be the same [33].

Feature extraction is a primary operation in image processing, that is to say, it is the first arithmetic processing of an image [34]. It checks each pixel to determine whether the pixel represents a feature. If it is part of a larger algorithm, the algorithm usually checks only the feature area of the image. As a precondition for feature extraction, input images are smoothed in scale space through Gauss fuzzy kernel. After that, one or more features of the image can be calculated by local derivative operation [35]. Sometimes, if feature extraction requires a lot of computing time, and the time that can be used is limited, a high-level algorithm can be used to control the feature extraction class, so only the part of the image is used to find the feature. Because many computer image algorithms use feature extraction as its primary computing step, a large number of feature extraction algorithms have been developed, and their extraction features are varied, and their computational complexity and repeatability are very different [36].

3.2. Characteristic Type. The edge is a pixel that forms the boundary between the two image regions. Generally, the shape of an edge can be arbitrary and may include intersection points. In practice, the edge is generally defined as a subset of points with large gradients in the image. Some commonly used algorithms will also link high gradient points to form a more perfect description of edges. These algorithms may also impose restrictions on edges. Locally, the edge is a one-dimensional structure.

Corner is a point-like feature in an image, and it has a two-dimensional structure in part. In the early algorithm, edge detection was first carried out; then, the trend of the edge was analysed to find the sudden turning angle of the edge. The algorithm developed later does not require the edge detection step. Instead, it can directly search for high curvature in the image gradient. Later, it was found that sometimes
a region with the same angular features could be found in the original image without corners.

Unlike angles, the region describes a regional structure in an image, but the region may also consist of only one pixel, so many regional tests can also be used to detect angles. A regional monitor detects a region that is too smooth for the angle monitor in the image. Area detection can be imagined as narrowing an image and then corner detection on a reduced image. Results of comparison test are shown in Figure 3.

Long striped objects are called ridges. In practice, the ridge can be regarded as a one-dimensional curve representing the symmetry axis, and the local needle has a ridge width for each ridge pixel. Extracting ridges from Gray gradient images is more difficult than extracting edges, corners, and regions. In aerial photography, ridges are often used to distinguish roads. In medical images, they are used to distinguish vessels.

3.3. Feature Extraction Step. Chi-square test: (1) the total number of documents in the statistical sample is $N$. (2) The frequency of positive documents ($A$), the frequency of negative documents ($B$), the frequency of positive documents, and the frequency of negative documents are calculated. (3) Calculating the square value of each word. (4) Each word is sorted according to the chi-square value from large to small, and the first k word is taken as the feature, and $K$ is the feature dimension.

Information gain: (1) the number of documents in statistical positive and negative classifications: $N_1$ and $N_2$. (2) The frequency of positive documents ($A$), the frequency of negative documents ($B$), the frequency of positive documents, and the frequency of negative documents are calculated. (3) Computational information entropy. (4) Calculating the information gain of each word. (5) Each word is sorted according to the information gain value from large to small, and the first k word is taken as the feature, and $K$ is the feature dimension.

3.4. Text Feature Extraction. Many machine learning problems involve Natural Language Processing (NLP) and must deal with text information. Text must be transformed into quantifiable feature vectors. Next, we will introduce the most commonly used method of text representation: the Bag-of-words model.

Thesaurus model is the most commonly used method of text modelling. For a document, ignoring its word order, grammar, and syntax, it is only regarded as a collection of words or a combination of words. The appearance of each word in the document is independent, does not depend on the appearance of other words, or when the author of the article chooses a word in any place is independent of the influence of the preceding sentence. The lexicon model can be regarded as an extension of the single thermal encoding, which sets an eigenvalue for each word. The lexicon model is based on similar words. Thesaurus model can achieve effective document classification and retrieval through limited coding information.

4. The Construction of Interactive English Chinese Translation Model

4.1. Introduction of Feature Extraction Algorithm. In this paper, the feature extraction algorithm is introduced, and the mapping of the best context is extracted to the translation process by the feature extraction algorithm, and the standard extraction of the feature context is completed. The optimal context is described by the semantic ontology mapping model. In the process of translation, there are $N$ translation contexts, including $K$ class semantic translation, and the number of translation contexts is $N_i$ ($i=1, 2, \ldots, K$). The probability of $K$ class semantic translation is $X_i = \{X_{i1}, X_{i2}, \ldots, X_{iN}\}$, $X_{ij} = \{i=1, 2, \ldots, K; j=1, 2, \ldots, Ni\}$; it is a directional $n$-dimensional vector result. Translation can be achieved by defining the process and translating the context into formula (1).

$$\alpha_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$$
In the formula, the alpha I is able to translate into the semantic translation context, and the best context alpha selection process is formula (2).

$$\alpha = \frac{1}{K} \sum_{i=1}^{N} \alpha_{ij}.$$  \hspace{1cm} (2)

4.2. The Design of the Language Processing Module of English Chinese Translation. Before conducting Interactive English Chinese translation, we first need to analyze the semantic transformation in English and Chinese translation and the word cover distortion components. In the process of English-Chinese translation, for example, as an example of the English word “image,” the English Chinese translation of the word may be a picture, but its synonyms are “image” and “picture.” Therefore, the three words have semantic fuzziness in a certain environment. In the process of English and Chinese translation, the semantic similarity and the possible semantic mapping of synonyms are described as formula (3).

$$\theta : S \rightarrow S \times [-0.5, 0.5].$$ \hspace{1cm} (3)

In the form of theta, theta is an approximate semantic in the process of English-Chinese translation; $S$ is a language
mapping in the process of English-Chinese translation; this article is an interactive English Chinese translation, and the degree of revision of semantic mapping is beyond the [-0.5,0.5]. In this paper, the mapping relation between English and Chinese translation semantics is used to solve the problem of selection and correction of ambiguity. Comparison of feature extraction algorithms in different references is shown in Figure 4.

The high-order replacement expression of the English Chinese translation semantics is Mountain [0, T], and the translation semantics of the literal translation semantics in the English Chinese translation process is represented by the matching translation set after the conceptual analysis. In which T represents the context judgment set of interactive translation semantics. By definition above, the analytic point in English Chinese translation sentence can be expressed by function Delta.

\[
\Delta : [0, T] \to S \times [-0.5, 0.5]. \tag{4}
\]

We use interactive English Chinese translation statements to translate the content two times, such as formula (5).

\[
\Delta(\beta) = \begin{cases}
S_E, & K = \text{round}(\beta), \\
\alpha_E = \beta - K, & \alpha_E \in [-0.5, 0.5).
\end{cases} \tag{5}
\]

In the form of \(\alpha_E\), alpha E is the substitution expression coefficient in English Chinese translation; \(S_E\) is an interactive semantic substitution of information code, and round is a language information integration operator in the process of English Chinese Translation.

4.3. Result Analysis. Figure 2 is an interactive English Chinese translation system based on feature extraction algorithm and a traditional interactive English Chinese translation system. In the process of translation, the number of points in the control points and the control chart on the left side of the translation process are shown. On the right, the traditional Chinese English-Chinese translation system has a controlled distribution.

The node control distribution can reflect the semantic and contextual relevance of the translation system. The distribution is loosely illustrated, but the translation is correct but lack of contextual coherence. Performance comparison of feature extraction algorithms under different standards is shown in Figure 5.

By analysing Figure 5, it can be concluded that the interactive English-Chinese translation system based on the feature extraction algorithm designed in this paper has achieved good test results in other algorithms and tests under different standards. Therefore, we believe that the establishment of this model is highly efficient and can provide some reference and reference for English-Chinese translators.

5. Conclusions

This paper designs an interactive English Chinese translation system based on feature extraction algorithm. The feature extraction algorithm is introduced to select the feature semantics and establish the semantic ontology mapping model. The optimal solution of the interactive English Chinese translation process is selected, and the English Chinese translation process is finally realized through coding. It is hoped that this study will enable accurate translation in the process of English Chinese translation.

Data Availability

All the data can be obtained from the author.

Conflicts of Interest

The authors declare that they have no competing interests.

Acknowledgments

Key Research Projects of Humanities and Social Sciences in Anhui Universities in 2017 “A Study of Professional Development of EOP Teachers Based on Students’ Occupational Competence” (No. SK2017A0735).

References

[1] S. Desroches, A. Lapointe, J. Marin et al., “Environmental Scan on Canadian Interactive Knowledge Translation Tools to Prevent Diabetes Complications in Patients With Diabetes,” Canadian Journal of Diabetes, vol. 45, no. 2, pp. 97–104.e2, 2021.
[2] F. Ragep, M. Jamil, and F. Tarro, “On astronomia: an arabic critical edition and english translation of epistle 3,” Journal for the History of Astronomy, vol. 5, 2016.
[3] M. Forcada, F. J. Ragep, and T. Mimura, “Astronomy in tenth-century Islamic encyclopaedism,” Journal for the History of Astronomy, vol. 47, no. 3, pp. 346–347, 2016.
[4] D. Raynaud, S. Gessner, and B. Mota, “Andaló di negro’s de compositione astrolabii: a critical edition with english translation and notes,” Archive for History of Exact Sciences, vol. 73, no. 6, pp. 551–617, 2019.
[5] P. Kormanik, “Using Google Translate (Tm) for Non-native English Translation of Breast Cancer Treatment Summaries,” Oncology Nursing Forum, vol. 43, 2016.
[6] A. Tzouma and L. C. Triarhou, “The cerebellar system and what it signifies from a biological perspective, by professor christofredo jakob (english translation),” The Cerebellum, vol. 15, no. 4, pp. 395–416, 2016.
[7] S. Bi, “Intelligent system for English translation using automated knowledge base,” Journal of Intelligent & Fuzzy Systems, vol. 39, no. 4, pp. 5057–5066, 2020.
[8] Q. Guan, H. An, X. Gao, S. Huang, and H. Li, “Estimating potential trade links in the international crude oil trade: A link prediction approach,” Energy, vol. 102, no. 1, pp. 406–415, 2016.
[9] C. Y. Read, C. E. Ricciardi, A. Gruhl, L. Williams, and K. M. Vandiver, “Building genetic competence through partnerships and interactive models,” Journal of Nursing Education, vol. 55, no. 5, pp. 300–303, 2016.
[10] O. Texler, D. Futschik, M. Kučera et al., “Interactive video stylization using few-shot patch-based training,” ACM Transactions on Graphics, vol. 39, no. 4, 2020.
[11] J. Smith, I. Pike, M. Brussoni et al., “Mixed methods study exploring parent engagement in child health research in British Columbia,” BMJ Open, vol. 9, no. 5, article e025404, 2019.

[12] C. Valmohammadi and S. Dashti, “Using interpretive structural modeling and fuzzy analytical process to identify and prioritize the interactive barriers of e-commerce implementation,” Information & Management, vol. 53, no. 2, pp. 157–168, 2016.

[13] A. Tsimicalis, S. Le May, J. Stinson et al., “Linguistic validation of an interactive communication tool to help French-speaking children express their cancer symptoms,” Journal of Pediatric Oncology Nursing, vol. 34, pp. 98–105, 2016.

[14] J. Reimand, T. Arak, P. Adler et al., “Robust lane-detection algorithm based on optimal feature extraction, long short term memory neural network and error correction strategy,” Applied Energy, vol. 230, pp. 429–443, 2018.

[15] A. Shook and V. Marian, “The influence of native-language tones on lexical access in the second language,” Journal of the Acoustical Society of America, vol. 139, no. 6, pp. 3102–3109, 2016.

[16] U. R. Charrodiere, D. Rittenschober, V. Nowak, C. Nicodemi, P. Bruggeling, and C. Petracchi, “Fao/infoods e-learning course on food composition data,” Food Chemistry, vol. 193, pp. 6–11, 2016.

[17] X. Chen, K. Naehrer, and T. J. Applegate, “Interactive effects of dietary protein concentration and aflatoxin b1 on performance, nutrient digestibility, and gut health in broiler chicks,” Poultry Science, vol. 95, no. 6, pp. 1312–1325, 2016.

[18] M. Weniger, F. Kapp, and P. Friederichs, “Spatial verification using wavelet transforms: a review,” Quarterly Journal of the Royal Meteorological Society, vol. 143, no. 702, pp. 120–136, 2017.

[19] L. Kane and P. Khanna, “Vision-based mid-air unistroke character input using polar signatures,” IEEE Transactions on Human-Machine Systems, vol. 47, no. 6, pp. 1077–1088, 2017.

[20] C. Poncelet and F. Jacquemard, “Model-based testing for building reliable realtime interactive music systems,” Science of Computer Programming, vol. 132, pp. 143–172, 2016.

[21] N. Wang, D. Mo, Z. Song, R. Wang, and Y. Wu, “Feature extraction algorithm of a precession target based on a doppler frequency profile of dual-aspect observation,” Applied Optics, vol. 58, no. 17, pp. 4695–4700, 2019.

[22] J. Xu and Z.-H. Mao, “Multilabel Feature Extraction Algorithm Via Maximizing Approximated and Symmetrized Normalized Cross-Covariance Operator,” IEEE Transactions on Cybernetics, vol. 11, pp. 1–14, 2019.

[23] G. P. Wang, J. X. Yang, and R. Li, “UfKlda: an unsupervised feature extraction algorithm for anomaly detection under cloud environment,” ETRI Journal, vol. 41, no. 5, pp. 684–695, 2019.

[24] U. Ozgunalp, “Robust lane-detection algorithm based on improved symmetrical local threshold for feature extraction and inverse perspective mapping,” IET Image Processing, vol. 13, no. 6, pp. 975–982, 2019.

[25] A. J. Heitmann and R. S. Gardiner-Garden, “A robust feature extraction and parameterized fitting algorithm for bottom-side oblique and vertical incidence ionograms,” Radio Science, vol. 54, no. 1, pp. 115–134, 2019.

[26] H. O. Awa, O. U. Ojiabo, and B. C. Emecheta, “Integrating TAM, TPB and TOE frameworks and expanding their characteristic constructs for e-commerce adoption by SMEs,” Journal of Science & Technology Policy Management, vol. 6, no. 1, pp. 76–94, 2015.