International Journal of Science and Engineering Applications
Volume 12-Issue 02, 80 – 82, 2023, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1202.1029

English-Chinese Corpus Collection and Artificial Intelligence Translation Based on Dynamic Clustering Model

Lijuan Zhou
Yuzhang Normal University
Nanchang, Jiangxi, China,330103

Abstract:This paper studies the English-Chinese corpus collection system and artificial intelligence translation function based on the dynamic clustering model. This paper attempts to introduce the semantic analysis of English-Chinese parallel corpora into the semantic recognition system. By labeling different levels of the corpus, including grammatical labeling, shallow semantic labeling and deep semantic labeling, the corresponding English-Chinese translation teaching system is designed, and the teaching system is improved by using the comparatively systematic teaching effects in actual teaching. On the basis of self-built Chinese-English parallel corpus, this paper extracts the Chinese high-frequency verb "to improve" to discover and summarize its English translation rules in different texts.

Keywords: English-Chinese Corpus, Artificial Intelligence Translation, Dynamic, Clustering Model

1. INTRODUCTION

The list of commonly used words in modern Chinese (draft) ranks "you" in the sixth place in common Chinese words; in the "Chinese Word Frequency Dictionary" by Xiao (2009), "you" is ranked as "character" in the ninth place. "Words" ranked tenth. This dictionary is based on statistical analysis of Chinese corpus of 73 million Chinese characters or 50 million words, and these corpora include four major registers: spoken language, news, fiction, and non-fiction. Therefore, we can conclude that the verb "you" is indeed a common verb in Chinese. As a commonly used verb in Chinese, it appears so frequently in Chinese, how to translate "Yes" according to the context in the translation process? Is it translated into "have", "has" or "there be" whenever you see "you"? [1-6]

With the rapid development, popularization and globalization of the Internet, on the one hand, the proportion of non-English texts in online text resources has increased rapidly. In the field of news media, more and more news portals provide multilingual news. Intuitively, the collocation between words has a certain rule, which is manifested as a relatively fixed collocation relationship between some words and other words. With the development of big data technology, machine translation has entered the stage of intelligent translation of neural network, with the breakthrough and development of corresponding artificial intelligence technology (Artificial Intelligence, AI). The BBC, the UK's largest news broadcaster, provides broadcasts in 43 languages and reports from 27 news networks. China Daily, Xinhuanet, China News, Tencent News International Channel, etc. [7-14].

Of course, this method of authentic language materials has obvious advantages, that is, it excludes the inauthenticity of man-made languages. Hunston (2002) even said: The application of language research has undergone a revolution". In terms of bilingual parallel corpus research, the history of European research is relatively long. It is generally believed that the article published by MonaBaker (1993) is the first chapter in driving the translation of corpus research; Wang Kefei of Beijing University of Foreign Studies has conducted large-scale research in China. 2003), carried out relevant research on the English-Chinese and Chinese-English parallel corpus of 30 million words that he established. all provide news in English and Chinese. For example, by observing the collocation of the words "obvious", "further", "improved", and "improved", it is not difficult to find that either "significantly improved" or "significantly improved" can be said. Machine translation based on artificial intelligence has become a hot topic in recent years. Using artificial intelligence technology, machine translation errors can be reduced by about 60%, improving the accuracy of translation. "Hunan Bilingual Network", a subsidiary of China Daily, began to provide bilingual news reading. Some major international companies and well-known institutions have gradually begun to provide multilingual versions of their website information. [15-21].

Either "further improvement" or "further improvement". In addition, "obvious" and "further" are both adverbs that characterize the degree, while "improvement" and "improvement" are both verbs that characterize the tendency to change. There is a relatively fixed collocation relationship between these two types of words. Companies such as Microsoft, Google, Baidu, and iFLYTEK have launched artificial intelligence-based online translation systems. Driven by information technology, many language service companies have also proposed development ideas based on big data and artificial intelligence. These changes have resulted in the rapid accumulation of multilingual text resources. On the other hand, countries around the world have a more urgent need for multilingual text organization and mining. Under the trend of globalization, a single language cannot meet the needs of a country or institution on the international stage. It can be considered that the relationship between words can be roughly divided into two types: semantic similarity and semantic correlation. For example, "obvious" and "further" are semantically similar, while "obvious" and "improvement" are semantically related. Given any two words, find their respective sets of semantically related words. [22-24].

2. THE PROPOSED METHODOLOGY

2.1 The Dynamic Clustering Model

Although machine translation based on neural network has made great progress, machine translation still lacks a lot in terms of deep semantic structure, different stylistic styles, language styles and discourse levels. Multilingual information organization and processing is an unavoidable problem. For non-English-speaking countries, it is not only necessary to deal with national language resources, but also a large number of resources in English. , according to the intersection of these two sets, it can be concluded whether the two words are semantically similar. This algorithm is based on the above ideas, and the large-scale annotated corpus [3] just provides resources for finding semantically related word sets. This means that the construction of corpus is still a very important foundation for the development of the artificial intelligence translation. By treating the corpus as a knowledge base, let the computer learn various knowledge from it, or treat the corpus as infinite. In recent years, related researches have proposed many text clustering techniques and their improved algorithms, but most of the researches are aimed at monolingual texts.

Donald Hindle of Bell Labs proposed a method for English noun clustering in his paper. In English, the verb-object or subject-verb collocation of verbs and nouns has certain rules. After years of development, my country has formed a number of corpora such as the "China-English-Chinese Parallel Corpus" and the China Language Resources Alliance, which have been widely used in translation teaching

2.2 The English And Chinese Corpus Collection

English and Chinese news texts, and then merged the clusters. When calculating the similarity between clusters, they used the method of translating the nouns, named entities and verbs in Chinese news into English. . The specific method is that, for any given noun, a set of a series of matching verbs can be obtained in the corpus. Each verb in the set will have a co-occurrence probability with the noun, so it has a certain amount of interactive information. , the calculation of the amount of interactive information is based on the formula.

The corpus integrates structure, semantics, contextual variables, and language typology attributes based on linked data models by fully recording language structure and functional characteristics. The method of first clustering and then merging cannot consider the whole text, and does not consider the importance of feature words from the perspective of all texts. Wen Yang et al. from the Intelligent Technology and System Laboratory of the Department of Computer Science, Tsinghua University proposed a Chinese adjective-noun clustering method based on collocation pairs. In Chinese, nouns and adjectives have a relatively fixed collocation relationship. We have entered the era of 3.0, but what are the characteristics of the constituent elements of the corpus under big data, and what are the characteristics of the relationship between the corpus and artificial intelligence translation? Lawrence did a Russian-English multilingual text clustering study using machine translation systems and dictionary word-by-word translation in 2003.

Taking nouns as entities and adjectives as features, the clustering of nouns can be obtained. Similarly, taking adjectives as entities and nouns as features, the clusters of adjectives can also be obtained, but the clusters of different parts of speech are related to a certain extent. This paper aims to start with the analysis of the characteristics of big data corpus under artificial intelligence translation, put forward the construction ideas of big data corpus, and provide suggestions on how to integrate the reform of translation teaching into the era of big data artificial intelligence. This study selects English and Chinese bilingual news texts as the experimental corpus, based on the following considerations: compared with other types of texts, news texts contain a large number of information features, which can highlight the theme of the article; bilingual texts are compared with monolingual texts.

2.3 The Artificial Intelligence Translation

They cannot be separated. The clustering process of the two is interactive. Based on the above ideas, they proposed a bidirectional hierarchical clustering algorithm. By continuously alternating classification and clustering, the number of classes is reduced and the number of words in the class is increased. , and cluster the two types of words at the same time. In the Internet age, user-generated content is an important feature, and many valuable corpus data often come from community discussions, customer blogs, WeChat groups, etc. The source of corpus data can be formed by extracting and refining user-generated data. The development of modern technology has made the system more and more complex, and the corresponding simulation system has also greatly increased the complexity, which naturally has a greater impact on the M&S confidence, so the M&S confidence assessment will become more important.

The problem of confidence assessment still lacks effective means, among which the model verification method needs to be further studied. Bilingual dictionaries in general fields have inherent deficiencies in a specific field, and bilingual dictionaries in specialized fields are not updated in a timely manner. The use of experimental data and prior knowledge to test complex hypotheses with the help of Bayesian theory proposed in this paper is a method of making full use of experimental information, closely combined with simulation technology, and can carry out statistical inferences under very small samples. Notably, with the advent of machine translation and artificial intelligence, machine-generated data has emerged as a potentially viable data source. For example, Google Translate uses machine-generated data when testing its artificial intelligence system. Therefore, this study attempts to conduct clustering research on English-Chinese bilingual mixed texts without the need for professional dictionaries.

In addition, in Chinese paragraphs, English acronyms are often included. This algorithm requires some people's work, that is, people need to choose a few words that can better represent the class, but if there are better results, such work is worthwhile. The stock of data assets of core customers is actually huge. If only narrowly defined complete bilingual corresponding parallel corpus data may be less. This study selects English and Chinese bilingual news texts as the experimental corpus, based on the following considerations: Compared with other types of texts, news texts contain a large number of information features, which can highlight the theme of the article. In the word selection stage, we manually selected the verbs of 15 parts of speech. Obviously, this choice will have a certain impact on the results. If the selected words cannot represent the class well, the results obtained are unsatisfactory. , we can use other clustering methods. In the era of big data, pure machine translation is inseparable from the cooperation of experts, and the mode of human-machine collaboration will be an important part of the construction of big data corpus.

3. CONCLUSIONS

Based on the standard KMeans algorithm, this paper conducts a comparative experimental study of text clustering on English-Chinese bilingual corpus. Among the large number of English-Chinese control corpora, comparative experiments were conducted based on Chinese single language, English single language and English-Chinese mixed language. The close integration of simulation technology can make statistical inferences under very small samples, test complex hypotheses, complete model verification, and an effective way to improve simulation confidence.

4. REFERENCES

[1]Shang Wenbo. Corpus-based explicit characteristics of the logical relationship between English and Chinese academic translation texts: Taking Handbook of Social Justice in Education translation as an example [J]. Contemporary Foreign Language Studies, 2020(5).

[2] Lu Yan. The construction of big data corpus under artificial intelligence translation[J]. Gansu Science and Technology, 2019, v.35(17):86-90.

[3] Liu Jiaxiang. A corpus index construction method based on dynamic K-means algorithm:, CN110674243A[P]. 2020.

[4] Lu Jiawei, Wang Xiaoding, Gao Yanxu, et al. Fusion method of knowledge graph relation extraction and REST service visualization based on DBSCAN clustering algorithm:, CN111143479A[P]. 2020.

[5] Zhang Weiwei, Hu Yaqi, Zhai Guangyu, et al. Clustering method of academic abstracts based on LDA model and Doc2vec[J]. Computer Engineering and Applications, 2020, v.56; No.949(06):186-191.

[6] Chen Qiaoyun. A Corpus-based Study of English and Chinese Bright Words: A Cognitive Semantics Perspective[D]. Xiamen University, 2018.

[7] Wu Xian, Hu Junfeng. Design of online dictionary compilation system based on diachronic corpus[J]. Journal of Chinese Information Processing, 2020(5):9.

[8] He Mengxin, Liu Hongyun. Psychological Science Popularization Content Feature Mining: Based on K-means Algorithm and LDA Topic Model [C]// Abstract Collection of the 22nd National Conference on Psychology. 2019.

[9] Tang Lizhe. Research and Implementation of Distributed Topic Clustering Technology Oriented to Text Streams [D]. National University of Defense Technology, 2019.

[10] Wang Xing, Jiao Wenxiang, Tu Zhaopeng. Sentence translation method based on artificial intelligence, device: CN111553174A[P]. 2020.

[11] Yin Gongjun. Improvement of vector space model based on word vector[J]. Modern Computer, 2018, 000(036): 32-35,41.

[12] Deng Yaochen, Peng Weiming, Shen Minglei. A polysemous translation method based on artificial intelligence knowledge graph: CN108563643A[P]. 2018.

[13] Shi Yahui. The phenomenon of Chinese-English translation of Chinese names for common subjects——A corpus-based study[J]. 2021(2015-1):36-40.

[14] Liu Keqiang. Cluster analysis of the English translation of the Chinese common verb "you" based on the Chinese-English parallel corpus[J]. 2021(2010-3):33-35.

[15] Zhang Fang, Ma Guanghui. A case study on the Chinese translation of when an adverbial clause——Based on the English-Chinese corresponding corpus[J]. 2021(2013-1):124-128.

[16] Wang Kefei, Liu Dingjia. Investigation and analysis of Chinese translation of English passive structure based on a super-large English-Chinese parallel corpus[J]. 2021(2018-6):79-90.

[17] Zhen Fengchao. Exploring English-Chinese Translation Units from Valence Structure——A Corpus-based Investigation[J]. 2021(2016-3):442-454.

[18] Ma Jianjun, Zhu Mulangma, Liu Wenyu. A corpus-based study of manifestation in English-Chinese business translation[J]. 2021(2018-6):123-128.

[19] Wang Mengyao, Wang Xiaoye, Hong Ruiqi, et al. Evaluation object mining based on improved BIRCH clustering algorithm[J]. Software, 2019, 40(11): 5.

[20] Zhang Chi, Zhang Guanhong. Text clustering algorithm based on word vector and multi-feature semantic distance[J]. Journal of Chongqing University of Science and Technology: Natural Science Edition, 2019, 21(3): 5.

[21] Pu Wenlong, Peng Yuanyuan. A text corpus construction and optimization method based on BM25 algorithm: CN110134799A[P]. 2019.

[22] Qi Xianting. Research on K-means text clustering algorithm based on density peak optimization [D]. Wuhan University of Technology, 2018.

[23] Sun Zhaoying. Research on optimization of text clustering algorithm based on convolutional neural network [D]. Shanghai Jiaotong University, 2018.

[24] Xiao Qiaoxiang, Cao Buqing, Zhang Xiangping, et al. Web service clustering method based on Word2Vec and LDA topic model[J]. Journal of Central South University: Natural Science Edition, 2018, 49(12):7.