Journal of Artificial Intelligence Practice (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2022.050302
ISSN 2371-8412 Vol. 5 Num. 3

Application of Artificial Intelligence (AI) Technology in Chinese English Translation System Corpus

Jing Long*

Chengdu University, Chengdu, Sichuan, China
janelong2100@163.com
**Corresponding author*

Keywords: AI Technology, Chinese English Translation, Translation System, Corpus

Abstract: With the rapid development of the Internet, the state is vigorously supporting the Internet high-tech industry. The Chinese-English translation industry of AI belongs to the branch of Internet high-tech industry, which is in line with the development direction of national scientific research industrialization. Therefore, the Chinese-English translation corpus has become a popular direction of AI. The key content of Chinese-English translation lies in the construction of corpus. At present, most of the response corpora on the market apply to daily response, which lacks professionalism in a certain field. In view of the common shortcomings that the corpus construction on the market can not achieve the expected effect and the content lacks professionalism in a certain field, this paper proposes a hybrid intelligent Chinese-English translation corpus construction method. This hybrid intelligent corpus is different from the general daily question and answer corpus. This paper mainly discusses the construction method of corpus in Chinese-English translation system based on AI technology. The corpus of this translation system comes from manual collection. Considering the accuracy of response and the complexity of training, this paper writes the professional corpus into the situational response corpus based on AIML.

1. Background Significance

After entering the 21st century, under the background of economic globalization, the world has ushered in an era of endless scientific and technological innovation. The popularization of AI will not only bring model changes to the industry, but also make human life more comfortable and convenient. However, in the context of big data, due to the large data stored in the corpus, corpus tools can quickly retrieve and accurately analyze the corpus. Using corpus for translation research has become the mainstream trend. However, there are few British Chinese / Chinese French English corpora, and there are few applications of AI technology in Chinese-English translation system corpora. The research on translation in China is relatively late. Until the beginning of this century, domestic scholars began to study the phenomenon of manifest (implicit) and translation in English-Chinese translation from synchronic and diachronic categories on the basis of foreign research. The explicit (implicit) features of text presentation are systematically studied [1].

After years of research on the application of AI technology in Chinese-English translation

system corpus, researchers at home and abroad have made some remarkable achievements in the hot field of AI research. For example, Dubey P proposes that the first and most important step in developing any machine translation system is to study and analyze the language used. A comparative study of machine translation is carried out. This study helps to select translation methods, preprocess the source text, formulate inflection analysis rules, and other activities required to deal with some special situations related to dogri. From the relevant studies, we can see that the study of English and Chinese structures is a hot topic in different linguistic fields. It makes a comparison and analysis, and summarizes the definition of middle verb structure in English and Chinese. Secondly, it summarizes the constraints and characteristics of English and Chinese interlanguage. Thirdly, it classifies and summarizes the previous studies from different perspectives, including syntax and semantics, generative linguistics and cognitive linguistics [2, 3]. Fourth, according to the research between comparative English and Chinese intermediate structures, it has turned to recent years. Corpus based research mainly focuses on the intermediate structures of high-frequency words in Chinese. It analyzes the universality of intermediate structure in English and English. Chinese intermediate structure exists in different languages. They have different ideas and different language implementations. According to the above, it may be a summary of the recognition standards of English and Chinese middle constructions. As for the English middle verb structure, it can first be described as "NP + VP + adv". It is the syntactic realization of English Intermediate semantics. Second, NP is not an agent, VP is an active form, and adv is a necessary component of English middle verb structure. Another agent was not present at the intermediate structure. Third, ADV is usually an adverb after the predicate. Last but not least, “NP + VP + adv.” should conform to the customer's way, agency and responsibility, non-agent subject and genericity. As for China's intermediate structure, it can be described as "NP + v-qilai + AP". NP is not an agent. There is an implicit motivation in Chinese middle verb structure [4]. The semantics of this AP can point to "v-qilai" or implied agent. AP is an essential instrumental component, usually an adjective in Chinese intermediate structure. Different scholars choose different perspectives to analyze English and Chinese intermediate structures. Previous studies have mainly focused on the definition and the constituent features of English and Chinese middle constructions. There is little research on the operating mechanism of English Chinese intermediate structure, which needs further research.

2. Chinese English Translation System Corpus

Corpus is a collection of electronic texts composed of specific sampling standards, which can represent a language, a variant or a language type. There are several corpus types; For example, the first category is general reference corpus and special purpose corpus; The second is oral corpus and written corpus; The third is bilingual parallel corpora, that is, multilingual corpora, parallel corpora and monolingual corpora; The fourth is learner corpus. Corpus linguistics is the basic means of language research, which emphasizes taking natural language as the research object, including oral and written texts. Corpus is not the purpose and object of research, but a method or means of research. Corpus technology mainly includes vocabulary index, thesaurus, collocation, keywords and so on.

2.1. Composition of Chinese English Translation System

(1) In Chinese English translation system, there are generally three parts: NP, VP and AP. Its commonness lies in: NP can be a patient or a locative noun; Pronouns can be used in the subject position to refer to nouns that have appeared before; VP is the main dynamic in form, and the verb is the transitive verb that represents the action, but it is inferior to the materiality in the middle construction; AP can express the intrinsic properties of NP. The middle construction implies an

uncertain agent, that is, no matter who is the agent, the meaning of the middle construction will not be affected. The main differences between Chinese and English middle constructions are: in Chinese middle constructions, if NP has appeared in the preceding text, NP can be allowed to be vacant; In English middle construction, NP cannot be omitted. Verbs in Chinese are action verbs, but they often need to be combined with grammaticalized "get up", while verbs in English can be completion verbs. AP in Chinese middle construction is usually an adjective indicating difficulty or nature, while AP in English middle construction is usually an adverb indicating value or way [5].

(2) From the perspective of construction grammar, the relationship between the three components of Chinese English Chinese construction is mutually restrictive. The commonness is that the attributes of NP can be reflected in its physical role, which has selective restriction on the VP entering the middle structure; The semantics of AP can point to the attributes of NP. The main differences are: AP in English verb construction is often an adverb to modify VP; In Chinese middle construction, as an adjective of AP, its semantics can point to the action behavior represented by "VP" or "VP + NP".

(3) Grammatical metonymy plays an important role in the operation mechanism of middle construction. However, the metonymic cognitive models of Chinese and English verb constructions are different. The metonymic model of verb construction in English is "action generation result" [6, 7]. In Chinese middle construction, metonymy model is "action instead of action". Its cognitive motivation is the prominence of non agent role and the weakening of agent role.

2.2. Experimental Verification Environment of Situational Intelligent Corpus Construction Method

In order to verify the corpus construction method of translation system described in this paper, AI technology is constructed as the evidence environment. The whole process can be summarized as that the user puts forward text or voice. If it is voice, the system will extract keywords from the other party's questions, and the hybrid intelligent translation corpus will output the English answers corresponding to the input Chinese. At the same time, the exchange database constructed in this paper is to save the dialogue content for later corpus training [8].

2.3. AI Chinese English Translation Corpus Preprocessing

In this paper, the design of Chinese-English translation corpus preprocessing is divided into two parts: Chinese word segmentation and word vector generation. Spaces are generally used for word segmentation between English words, but there is no separator between Chinese words. The computer's understanding of natural language semantics depends on the accuracy of word segmentation. On the basis of meeting the accuracy of word segmentation, the faster the speed of word segmentation, the faster the response of intelligent response corpus. In view of the above analysis, this paper needs to meet the two basic requirements of word segmentation accuracy and speed when considering word segmentation [9, 10].

Chinese word segmentation refers to the segmentation of a series of Chinese characters into individual words. There are many kinds of word segmentation tools in Python, including Yaha word segmentation, Pangu word segmentation, stuttering word segmentation, etc. The grammars of these word segmentation tools are the same, and Yaha word segmentation tools cannot deal with words such as "yellow glazed tile roof" or "round Mound Altar". Through the comparison of actual operation, it is found that the stuttering word segmentation tool is the most suitable tool for word segmentation. This paper finally selects the stuttering word segmentation tool with accurate mode, full mode and search engine mode[11].

3. Algorithm for Intelligent Mining of Massive Digital Archives

That is, words are expressed as vectors. As the simplest word vector representation method, one. Hot representation has the disadvantages that the word vectors of any two words are orthogonal, which can not reflect the semantic similarity between words and the thesaurus is too large, resulting in too high dimension. Compared with one. Hot representation, the former can consider the context information in the current context, better express the similarity between words, and provide more abundant word vectors with semantic information.

3.1. Cbow text representation model

Cbow (continuous bag of words) text representation model predicts the central word through context. Cbow text representation model is trained to find the maximum value of objective function Z. the likelihood function formula of logarithm for objective function optimization is shown in (1):

$$Z = \sum_{s \in q} \log q(S / content(S)) \tag{1}$$

Input layer: for 2q word vectors s, use Vr-q, Vr-q+1…Vr-1+q, Vr+q.

Projection layer: sum and accumulate 2q word vectors in the input layer;

$$Y_s = SUM(S_{r-q}, S_{r-p+1}, ...S_{r-1+q}, S_{r+q}) \tag{2}$$

Output layer: the output layer of cbow text representation model uses the number of occurrences of each word in the input corpus as the weight to construct the Huffman tree;

4. Investigation and Research Process and Analysis

The interface design of AIML intelligent response corpus is to use PHP to design the constructed AIML response corpus into the form of API interface for PC or mobile client call, and present satisfactory results to users. That is, when the user asks a question, the interface can return a text in JSON format. After being parsed by JSON in Android development, it can be converted into our common string, and finally presented to the user through Android client.

The interface API of AIML response corpus adopts PHP and is developed based on netbeanslde development tool. Detailed design process of interface table 1:

Table 1: Detailed design of interface

Interface address	URI		method
Request parameters	Api. php		POST
	parameters		value
	Input		xxxx
Return parameters	parameters	type	remarks
	status	Int	201
	error	String	success
	requestType	String	tall
	result	String	Output dialog

In addition, by analyzing the correspondence, manifestation and implicitness of logical connectives in Chinese-English translation, is it mainly based on turning relationship, conditional relationship, causality, or Skopos relationship? Are there any differences between the four? Detailed data sorting is shown in Table 2 and Figure 1:

Table 2: Logical relations in Chinese English translation

Logical conjunction type	Corresponding		Quantification		Manifest	
	Quantity	Frequency	Quantity	Frequency	Quantity	Frequency
Turning relationship	245	51.04%	98	56.65%	67	29.13%
Conditional relation	54	11.25%	34	19.65%	8	3.48%
Causal relationship	162	33.75%	34	19.65%	76	33.04%
Purpose relationship	19	3.95%	7	0.06%	79	34.35%
Total	480	100%	173	100%	230	100%

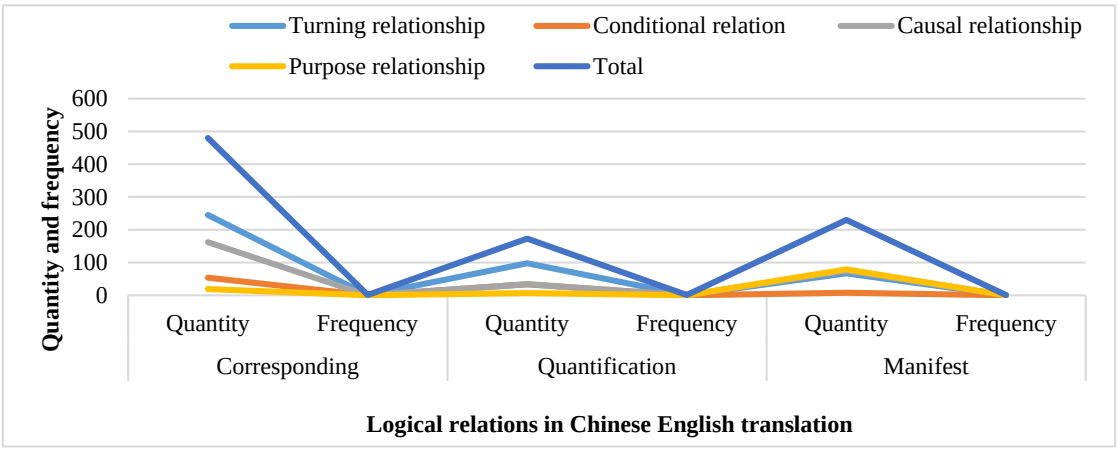

Figure 1: Logical relations in Chinese English translation

5. Conclusions

In recent years, driven by big data technology, corpus based explicit research has developed vigorously. The Chinese English translation system is an important way for the international community to understand China's politics, economy, diplomacy, society and culture, and plays an important role in strengthening the construction of international discourse power. Compared with the study of literary translation, the study of Chinese-English translation in China is not deep enough. Corpus based quantitative studies also mostly focus on English versions, the study of Chinese-English translation is one of the directions of corpus research.

References

[1] Dubey P. The Hindi to Dogri machine translation system: grammatical perspective [J]. International Journal of Information Technology, 2019, 11(1):171-182.
[2] Chen W L., Lin Y B., Ng F L., et al. RiceTalk: Rice Blast Detection Using Internet of Things and AI Technologies[J]. IEEE Internet of Things Journal, 2020, 7(2):1001-1010.
[3] Majumdar B., Sarode S C., Sarode G S., et al. Technology: AI [J]. British dental journal, 2018, 224(12):916-916.
[4] Leushacke M., Tan S H., Wong A., et al. Lgr5-expressing chief cells drive epithelial regeneration and cancer in the oxyntic stomach [J]. Nature Cell Biology, 2017, 19(7):774-786.
[5] Horii H. Advancement of Vehicle Occupant Restraint System Design by Integration of AI Technologies [J]. International Journal of Transport Development and Integration, 2021, 5(3):242-253.
[6] Anan T., Higuchi H., Hamada N. New AI technology improving fuel efficiency and reducing CO2 emissions of ships through use of operational big data [J]. Fujitsu entific & Technical Journal, 2017, 53(6):23-28.
[7] Raknys A V., Gudelis D., Guogis A. The Analysis of Opportunities of the Application of Big Data and AI Technologies in Public Governance and Social Policy [J]. Socialinė Teorija Empirija Politika ir Praktika, 2021, 22(6):88-100.
[8] Cui L. A Preliminary Study on the Management Strategy of University Personnel Files based on AI Technology [J]. Journal of Electronic Research and Application, 2021, 5(2):1-4.

[9] Xu K., Wang Z., Zhou Z., et al. Design of industrial internet of things system based on machine learning and AI technology [J]. Journal of Intelligent and Fuzzy Systems, 2021, 40(2):2601-2611.
[10] Jho, Gook-Hyung, Yoon, et al. French Tagsets for Developing Common Transfer Knowledge, that constitutes a Multilingual Automatic Translation System [J]. Cogito, 2017(82):337-375.
[11] Rui Z., Mao K. Topic-Aware Deep Compositional Models for Sentence Classification [J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2017, 25(2):248-260.