Construction of Corpus in Artificial Intelligence Age

Changchun Hong[1,a]

[1]School of Foreign Languages, Huangshan University, 39 Xihai Road, Tunxi District, Anhui Province, China

Abstract: As a kind of revolutionary technology, artificial intelligence marked an explosive transformation in many fields of study. Nowadays, much of the translation work used to be done by human has been undertaken by machines. The construction of corpus is a crucial step leading to successful machine translation. The paper aims to exploring the mode of corpus construction from the perspective of information mining, information retrieval and information processing. The retrieval system uses web crawlers to collect network information and automatic tagging technology to index the collected information, then applies corresponding language processing techniques to achieve correspondence between two languages and form an index database. In the age of artificial intelligence, machines can keep a track of many users' searches, queries, so as to record, extract as well as to feed back on different translations to build a new corpus. In this way, machine translation is improving in its scope and accuracy in translation with the goal to take up the tedious work of human translation as well as to increase the speed and reduce the cost of it.

1. Introduction

On April 8, 2018, the Boao Forum for Asia was held in China. In terms of conference translation services, translators and simultaneous interpretation equipment in the era of artificial intelligence have made an eye-catching appearance. After 60 years of development, artificial intelligence is currently undergoing the third wave. Its characteristics are based on new advances in the field of neural networks, using big data and cloud computing backstage as computing platforms, and based on mobile Internet sources, various training data are continuously received in the background, and feature data mining and rapid training are performed through deep learning capabilities.[1] The "machine simultaneous interpretation" technology used for the first time in this conference is actually one of the forms of artificial intelligence entering the formal use phase.

In recent years, major breakthroughs have been made in artificial intelligence technology. The famous British scientific journal "Nature" summarized ten breakthroughs in the field of technology in 2016 and artificial intelligence is at the top of the list. Apart from Alpha Go's victory over the world championship of Go, it also features that machine translation with artificial intelligence reduces errors by approximately 60% (Nature, 2016). The international technology giants, such as Google, Apple, Microsoft, IBM, Face book, etc., are also continuing to develop language products with artificial intelligence. Google launched NMT (Neural Machine Translation), a neural network machine translation system, which has greatly improved the quality of machine translation. The development of artificial intelligence technology is beyond our imagination. Machine translation is improving the processing of information to make up for the lack of human translation in terms of speed and cost.

2. The Development of Artificial Intelligence Translation

As a frontier discipline in the current development of science and technology, artificial intelligence is a comprehensive inter-discipline developed on the basis of computer science, cybernetics, information theory, neuropsychology, linguistics and other disciplines. It is a research on machine's imitating the human brain to engage in cognition, memory, learning, association, and other thinking activities, so that it can solve complex problems that humans cannot deal with, essentially mimicking the human brain on thinking. [2] At present, artificial intelligence technology has been applied in many aspects, and machine translation is one of the important applications. The process of machine translation can be roughly divided into three phases: source text analysis, transformation of source and target text, and translation generation. In the specific machine translation system, according to the purpose and requirements of different programs, transformation of source and target text phase and the source text analysis phase are combined, while the translation generation phase is independent, so a correlation analysis on independent generation system is established. In such system, the characteristics of the target language are taken into account when analyzing the original language,

Corresponding author e-mail: hongcc@hsu.edu.cn

while the features of the original language are not considered when the target language is generated. This system is generally used when translating a language into multiple languages. Some also separate the original text analysis phase, and combine the original text translation phase with the translation generation phase to establish an independent analysis-related generating system.

The system of machine translation has undergone continuous evolution and upgrading. One is a system based on rules, which has evolved from lexical, grammatical, semantic, and intellectual intelligence. Another type of system is based on corpora. [3] At present, the latest translation machine is an artificial neural network based on algorithm system whose core technology is a deep neural network with massive nodes (neurons) that can automatically learn translation knowledge from corpus. After a sentence in a language is vectorized, it is passed through the network and converted into a representation form that the computer can "understand", and then through multiple layers of complex conduction operations it can generate a translation in another language. As a result, a super-size corpus or even a hyper-size corpus is needed. Table 1 shows the differences between traditional corpus and AI corpus in capacity and time span.

Table 1. Comparison of corpora

English	Words	Language	Time period
Global Web-Based English(GloWbE)	1.9 billon	20 countries	2012-2013
Corpus of Contemporary American English(COCA)	450 million	American	1990-2012
Corpus of Historical American English(COHA)	400 million	American	1810-2009
TIME Magazine Corpus	100 million	American	1923-2006
Corpus of American Soap Operas	100 million	American	2001-2012
British National Corpus(BYU-BNC)	100 million	British	1980-1993
Strathy Corpus (Canada)	50 million	Canadian	1970-2000
Compare			
Google Books: American English	155 billion	American	1500s-2000s
Google Books: British English	34 billion	British	1500s-2000s
Google Books: One Million Books	89 billion	Am/Br	1500s-2000s

Note: Summarization of data from http://corpus.byu.edu/coca/

Google has always been a leader in the field of machine translation. With its powerful statistical method and cloud computing technology, the company's corpus provides instant translations among 65 major languages, including words, sentences, text, and web page translations.

In the mean time, the LSTM (Long Short-Term Memory) Recurrent Neural Network (RNN) is widely used in machine translation. The model is good at modeling natural language, transforming sentences of any length into floating-point vectors of specific dimensions, and at the same time "remembering" the more important words in the sentences, so that "memory" can be stored for a relatively long time.[4] The model solves the difficult problem of vectorization of natural language sentences, and it is of great significance to the use of computers to deal with natural language. It makes the computer's processing of language no longer stop at the simple level of matching, but further semantic understanding, which realized the translation method of "understanding language and generating translation". The biggest advantage of this algorithm lies in its smooth translation, which is more in line with grammatical specification, close to natural language and easy to understand. Compared with previous translations, there has been a qualitative leap, and in some contexts translations are almost the same as human translations. Google has always been a leader in machine translation, for it can provide instant translations between 65 major languages, including words, sentences, text and web pages, thanks to its powerful statistical methods and cloud computing technology.

3. The model of corpus construction

3.1 Information mining

At present, corpus linguistics mainly deals with the collection, storage, retrieval, statistics, grammar tagging, syntactic and semantic analysis of machine-readable natural language texts, and the application of corpus with the above functions in language teaching, quantitative analysis, vocabulary research, collocation research, lexicons and editing, grammar research, language and cultural studies, legal language research, work style analysis, natural language understanding, and machine translation. [5] The driving research methods are:

Concept definition →Database construction →Confirmation of working methods →Extraction of multiple word structures →Analysis on linguistic level (grammatical level, semantic level and pragmatic level)→Functional clustering (conceptual function, program function and interpersonal function).

The web intelligent mining technology in the era of artificial intelligence makes the expected collection more simple and convenient. Smart mining, also known as data mining or knowledge mining, is mainly a process of extracting from a large number of fuzzy and random practical application data that people do not know in advance but it belongs to potential useful information . The network intelligence mining technology needs to combine the network browsing and Internet load law research, network reverse engineering research and

network ontology engineering research. To excavate network data and knowledge, it is necessary to have a super-storage analysis and conversion function system, adopting multimedia data mining technology and network data warehousing technology to optimize network resources and further improve the utilization of network resources. Multimedia mining refers to the combination of data mining technology and multimedia information processing technology into information processing methods that help discover knowledge in multimedia data. The major sources of data objects for large-scale media data mining and processing include webcasting, digital libraries, internet TV, and internet newspapers. The forms of mining mainly include the comprehensive mining of images, graphics, text, audio, video, animation, Web, and multimedia. Mining structure consists of three parts: mining interface, multimedia database and multimedia mining engine. [6] In terms of user services, automatic recommendation strategies and different login pages are set according to user needs, and data analysis for long-term tracking of user journeys and mining of learner weblog information are implemented. In order to effectively improve the accuracy and reliability of text classification, intelligent mining of network texts has been introduced. Based on the dynamic clustering method and the classification of feature attributes, the new text data classification system of hybrid fuzzy clustering theory has been put into practice by CNKI.

3.2 The Retrieval of Computer Information

The widespread application of artificial intelligence technology in information retrieval systems is the result of the successful combination of artificial intelligence technology and information technology. In cross-language retrieval, the source language used in questioning is generally the user's native language, and the target language used in the retrieved document is generally a language that the user is not familiar with or even completely unfamiliar with. At present, the main methods for cross-language retrieval are question-based translation methods, document translation methods, question-document translation methods, intermediate translation methods, proper noun transliteration methods, and ontology-based conversion methods. The most commonly used is the question-based translation method. Corpus, especially the application of parallel corpora, not only improves the certainty of word translation, but also has important implications for the translation of proper nouns, because in the parallel corpora, the correspondence between words is unique, and many words in both the hand-made bilingual dictionaries and that cannot be obtained in the machine-readable dictionary can be obtained in the parallel corpora. The use of various language resources in cross-language retrieval is not isolated, and using two or more language resources simultaneously will achieve better results. To achieve the transformation between languages, computers must be made to understand the meaning of natural language texts, and then use natural language texts to express given intentions and ideas.

Cross-language information retrieval is based on the understanding of natural language, so its key issue is to make the query language and document language reach agreement before retrieval. If users ask questions in one language, it can retrieve relevant information described in another language or multiple languages. For example, if you enter a Chinese search query, the cross-language search system will return information in English, Japanese, and other languages. These information include not only text information, but also other forms of information.

The key technologies mainly involved in cross-language retrieval are computer information retrieval technology, machine translation technology and disambiguation technology. Information retrieval technology completes the matching between questions and documents and machine translation technology accomplishes semantic equivalence between different languages, while disambiguation technology solves the problem of polysemy and ambiguity in the translation process. Computer information retrieval technology is mainly about automatic search technology, automatic indexing technology, language processing technology and automatic matching technology.[7] The retrieval system uses web crawlers to collect network information and automatic tagging technology to index the collected information, then applies corresponding language processing techniques to achieve correspondence between the two languages and form an index database. The user enters a search formula, and the computer matches the search formula with the index item in the database, and exports the search result according to the size of the relevance degree between the search formula and the index item. Its implementation process is shown in Figure 1.

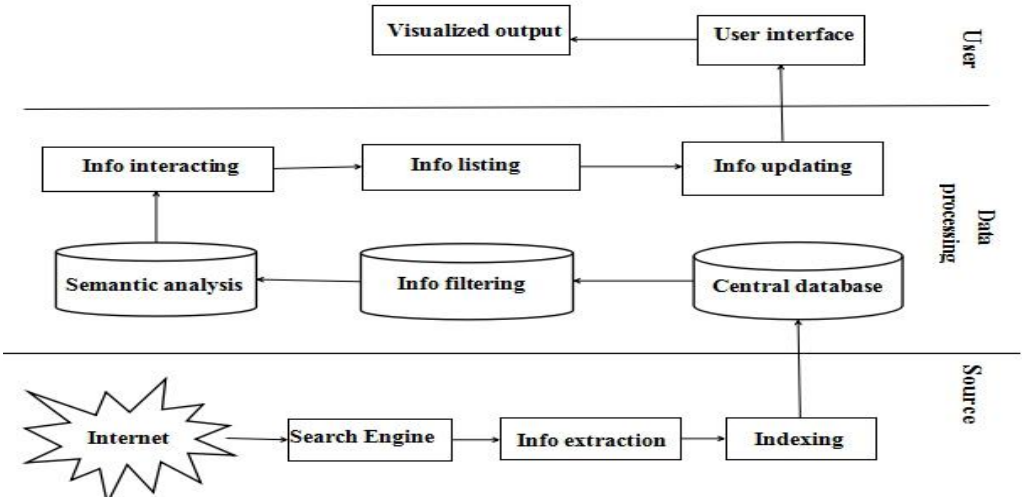

Fig. 1 Framework of information mining & processing

3.3 Machine Translation Technology

In cross-language retrieval, what needs to be solved is actually a language processing problem. Different from the monolingual language information retrieval and machine translation, it is not a simple superposition of two technologies, but an organic integration, with its own characteristics and special research content. Machine translation technology is essentially a computer program that can automatically translate the text of one language into another language, whose core is to maintain the semantic equivalence of two kinds of text (the source language text and the target language text).[8]

As the words in the source language text often correspond to several words described in the target language in the translation process, the most appropriate word should be selected or related processing should be applied to achieve the same meaning. In cross-language retrieval, the accuracy of translation directly determines the accuracy of the search. The improvement of accuracy requires the combination of natural language processing and machine translation. Since this involves sophisticated computer semantic analysis techniques, the disambiguation technique is the key to clear this obstacle.

Cross-language information retrieval involves the mutual conversion between the two languages. The main problem that arises in this process is the problem of ambiguity , which needs to address the wide variety of ambiguities that exist at all levels of natural language texts and dialogues. In natural language, the phenomenon of polysemy is very common. When dealing with queries, it is very important to determine the exact meaning of the search term, that is, to convert natural language input with potential ambiguity into an unambiguous computer internal indication, which requires a lot of knowledge and reasoning. For the searched documents to increase the precision rate, it is necessary to clarify the meaning of the search terms that appear in the literature to determine their relevance. The ambiguity in cross-language information retrieval comes from both the source language and the target language. The solution can mimic the human solution to the disparity to a certain degree, combining the grammar, lexical, syntax, semantics, etc., thus the analytical expression of text non-ambiguity can be achieved. However, it is very difficult for the machine to achieve correct and effective analysis at this full-text level.[9]

4. Conclusion

In the age of artificial intelligence, the machine can centralize the processing of many users' searches, queries, and feedback on different translations to build a new corpus, and artificial intelligence applies computer artificial intelligence theory to language understanding and natural language processing programs through procedures such as program editing, database processing, and data structure, which makes the perfect fusion of machine translation and artificial intelligence. However, in some areas, the level of machine translation is still far from the level of human translation, such as recognizing deep semantic structures, different literary styles and language styles. Even if it is synchronous machine translation, there is not much deviation from the perspective of the sentence. Therefore, the combination of corpus-based machine translation and manual translation can provide more efficient and quality translations. Human-machine translation, which uses corpus and artificial intelligence technology to do initial translations with high precision, and then the senior translators will further translate the literary and professional aspects of translated texts, and machine learning technology will also participate in this process. After learning, the machine will learn correct expressions and human language habits based on the results of human corrections, so as to optimize its corpus and upgrade the future translation ability. On the other hand, translation machines in the artificial intelligence age also have the ability to track the "hard" parts of human translation to prevent translators from making some low-level or logical mistakes based on its super-size corpus.

Acknowledgements

This paper is financially supported by the following :
Anhui Social Sciences Key Project Funding SK2016A0878
Shanghai Foreign Language Education Press hxkt20180001

References:

1. Foster I, Kesselman C, Nick J, et al. Computer Grid Services for Distributed System Integration[J]. IEEE Computer, 2002, 35(6):37-46.
2. Wang, Wenjie Principles and Applications of Artificial Intelligence. Beijing: People's Post and Telecommunications Press. 2004:25
3. McEnery , T .&Wilson, A.Corpus Linguistics [M] .Edinburgh University Press .1996.121-123
4. Greies. Corpus Linguistics and theoretical linguistics: A Love－hate relationship? Not necessarily［J］ International Journal of Corpus Linguistics，2010，15 (3) :327
5. Biber, Douglas Corpus linguistics [M].Cambridge: Cambridge University Press,2013:202
6. Tang, ZhaoHui. The application of data mining and processing [M]. Qinghua University publishing Press, 2007:97
7. Zang, Jingsong The application of artificial intelligence in information retrieval in inter-lingual contextualization [J]. Computer Era, 2016,10 :30-31.
8. Wray, Alison. Formulaic language: Pushing the boundaries［M］. Oxford: Oxford University Press，2008: 9.
9. Börkur S, Jaap K, Maarten R. EuroGOV: Engineering a Multilingual Web Corpus[C]//Proc. of the 6th Workshop on Cross Language Evaluation Forum. Berlin, Germany: [s. n.], 2006: 825-836.