Research on Translation Corpus Building with the Assistance of AI

Xiaoqing Hou*

College English Teaching Department, Zao Zhuang University, Zaozhuang, China
**Corresponding author. Email: 236913006@qq.com*

ABSTRACT

The quality of machine translation depends on the corpus available for it to learn from. With the development of AI, the translation corpus is now the key to realize Intelligent translation and scenario-based translation. The data on the Internet, data of the enterprise, user-generated data, machine-generated data, all of which can be the sources of AI-Assisted translation corpus data. Types of these data are diversified. There are term data, text data, unstructured data, and knowledge graph. AI-Assisted translation corpus construction can collect data in the following four modes: open-source data based on the third-party, corpus-sharing mode based on crowdsourcing translation, self-learning mode based on machine, collaborative mode based on human-machine cooperation. In the future, corpus-based translation teaching will be an important part of intelligent translation.

Keywords: *AI, Translation Corpus, Translation Teaching*

1. INTRODUCTION

With the development of AI, machine translation has become a hot topic in recent years. Assisted by AI, errors of machine translation can be reduced by 60%[1], which improves accuracy. Companies such as Microsoft, Google and Baidu have launched AI-based online translation systems. Many language service companies have also launched products based on big data and AI[2].

But in some areas, machine translation is much worse than human translation. For example, machine translation cannot accurately translate text when it involves deep semantic structures, different stylistic styles and language styles[3]. When it comes to sentence, there are no big mistakes in machine translation, but when it comes to discourse, there are still many deficiencies in machine translation. Language is not only used to express words and sentence meanings, but also to express emotions, culture and even artistic conception. Language connotation is so complex that translation is a kind of re-creation, which reflects the wisdom and creativity of human beings. As far as technology is concerned, there is no way for machines to be creative, especially in literary translation. Corpus is an important foundation for the AI translation. A corpus is a knowledge base, from which computers learn all kinds of knowledge. A corpus is also a sampling representation of infinite, continuously generated texts. The study of automatic language processing is the main method of machine translation. The machine translation, which based on self-study and huge corpora, is the cornerstone of the language services industry[4].

After years of development, China has built up several corpora, such as "Chinese-English Parallel Corpus" and "Language Resource Alliance of China", which are widely used in translation teaching, translation research and translation practice. Corpus translatology has formed and developed[5]. With large scale data, corpus can fully record the structure and functional characteristics of language. On the basis of associated data model, corpus integrates structure, semantics, context variables and language type attributes, and enters the 3.0 era[6]. However, what are the elements of AI-assisted corpus? What is the relationship between corpus and AI translation? How to build a corpus? These studies need to be further studied. Moreover, there is no parallel corpus built specifically for translation teaching[7]. Therefore, an important question for us to discuss is how to build an AI-assisted corpus, and we should combine the corpus with the translation practice and translation teaching together.

2. ELEMENTS OF AI-ASSISTED CORPUS

There are 3 elements of AI-Assisted corpus.

The first element is the neural network system, which is the core part of AI translation. Neural network system includes algorithm design, which changes the original way of machine translation, making machine translation from the rule-based and statistical stage to the stage of deep learning and reinforcement learning.

The second element is the application scenario. Nowadays, Internet technology has been applied to all fields of life, and there are more and more demands for scenario-based translation. For example, instant text, video and audio translation is needed in living and working scenes such as eating, traveling and seeing a doctor.

The third element is corpus. On the one hand, the quantity and quality of corpus data directly affect the effect of machine learning and the accuracy of translation results. On the other hand, the reading ability and specialization of corpus data directly support the realization of image, sound, video, location, trajectory and action. The multi-dimensional corpus data makes it possible to develop vertical artificial intelligence, which ensures intelligent translation and mobile translation.

Corpus data is so important that the key to the success of AI machine translation is how to combine algorithms and scenes to give full play to the value of corpus data.

3. CHARACTERISTICS OF AI-ASSISTED CORPUS DATA

Sources and types of AI-Assisted corpus data are very different from traditional corpus data.

3.1. Sources of AI-Assisted Corpora

3.1.1. Data on the Internet

A large amount of Internet data has strongly promoted the extraction and formation of corpus data. Extracting structured information from unstructured web pages is the main goal of building corpus data. For example, we can use the home page of a multilingual company to crawl and automatically align multilingual websites.

3.1.2. Data of the Enterprise

Data of the enterprise is an important and reliable source of corpus data. The operating instructions and data information in the enterprise database, are readily bilingual and aligned, with good quality and high completeness.

3.1.3. User-Generated Data.

An important feature of the Internet era is that the users generate content. Many valuable corpus data often come from community discussions, customer blogs, Wechat groups, etc. One of the sources of corpus data is to extract and refine user-generated data.

3.1.4. Machine-Generated Data

With the development of machine translation and AI, machine-generated data has become a potential and viable data source. Google Translate, for example, uses machine-generated data when testing its AI systems. For example, we can make machines generate similar sentences based on sentence examples, or we use the existed machine translation model to generate bilingual corpus by reverse translating monolingual corpus.

3.2. Types of AI- Assisted Corpus Data

AI-Assisted corpus data is multimodal and it is the integration of audio, video and text information[8-9]. The main types of AI-Assisted corpus data are as follows:

3.2.1. Term Data

Guided by professional knowledge, term data is an important basis for machine translation and play an important role in improving the accuracy of machine translation. There are massive term translation libraries in the Internet era. Term data here include dictionaries and other term libraries. Automatic term extraction is a hot topic in parallel corpus database construction[10]. However, the current online translation dictionaries are basically transformed from the paper version to the online version, which has not broken through the pattern of Chinese-foreign dictionary and has not formed the concept of multi-lingual term management[11]. As the terms are professional and authoritative, the construction of professional corpus database will be the focus.

3.2.2. Text Data

Textual data is still the first choice of corpus data. At present, the corpus of machine translation is still text-based, which contains tens of millions of corpus data. Text plays a very important role in machine translation. The rich semantic information contained in text can provide important support for the research and application of question-answering system, information retrieval, semantics, information extraction and other fields. Textual data is the most important part of ontology. The amount of textual data is huge, and it is easy to obtain, so machine translation based on textual data has become a key research object.

3.2.3. Unstructured Data

In the era of big data, unstructured data, such as voice, has become the main data source. Big data comes from web texts, web files, micro-blogs, videos and search engines. Compared with traditional text translation, unstructured and diversified data sources require more translation techniques. Corpus data can be formed through optical character recognition and automatic speech recognition. Language services are no longer limited to interpreting and translating documents. Translations based on mobile phone scans and mobile phone photos grew by nearly 10%. Unstructured data will become an important type of data.

3.2.4. Knowledge Graph

Through the data mining, extraction and integration, we connect and integrate previously unconnected data to form a unified and global knowledge base, then the knowledge graph is formed. The data of corpus is similar to knowledge graph. We connect massive fragmented and scattered big data in the form of knowledge graph to meet the immediate, dynamic, fragmented language needs.

In a word, construction of AI-Assisted corpus is confronted with problems such as diversity of data sources and uneven data quality. We need to discuss how to build AI-Assisted corpus from multiple aspects, and put forward the construction mode of corpus based on these discussions.

4. THE CONSTRUCTION MODE OF AI-ASSISTED CORPUS

4.1. Open-Source Data Based Model on the Third-Party

Online translation systems led by Google and Baidu, provide the basis for us to obtain parallel corpus data. We can construct parallel corpus on the basis of existing third-party databases, which provides a foundation for AI-Assisted translation learning. For example, we can find English translation versions of the same Chinese sentence in Google Translate and Baidu Translate. Based on the combination of the two versions, we can form our own translation corpus. This mode requires high computing power. Open-source data is often Internet-based and is available in the form of online translations. Its content is a multimodal combination including video and audio. The amount of open-source data is huge, so it has high requirements on grasping ability and algorithm.

4.2. Corpus-Sharing Model Based on Crowdsourcing Translation

The sharing of corpus database can solve the shortcoming of insufficient corpus for AI-Assisted machine translation. The sharing of corpus data among AI enterprises, language service enterprises and users can realize the integration of corpus resources. For example, corpus mall can provide a good platform for corpus demander and supplier. Corpus demander and supplier can enjoy corpus data service for free by providing corpus data. The mall has functions such as search, upload, download, account management and point purchase. The platform supports Chinese-English two-way search with fast retrieval speed. There are more than 73 million sentences in corpus and 1.5 billion words in total, which provides ways to expand corpus data resources.

Through crowdsourcing translation, translators and non-translators participate in translation activities together, which produces a large number of translation results. Crowdsourcing translation is essentially a collaborative translation that connects individual translators with machine translation. This means that both corpus demanders and corpus providers appear in crowdsourcing translation activities. Therefore, it is possible to build a corpus data sharing platform based on crowdsourcing translation.

4.3. Self-Learning Model Based on Machine

The machine translation based on neural network has the ability of self-learning. We take full advantage of the machine's ability to learn, and the quality of machine translation can be improved by making the machine learn in a loop over a certain period of time. It has been proved that the translation effect of the same corpus content can be significantly improved after the machine has learned many times by itself. It is critical to design the algorithm of self-learning mode based on machine. For example, AlphaGo's self-learning relies on repeated training of 30 million chess records. Periodic loop training is not a simple repetition, but through the design of effective algorithms, it continuously strengthens and improves the learning ability and memory ability of the computer.

4.4. Collaborative Model Based on Human-Machine Cooperation

The cooperation between machine translation and human translation is an important feature of translation in the new era[12]. In the era of big data, machine translation is inseparable from human-machine collaboration, and this model will be an important model for the construction of big data corpora. The sampling of corpus, the formulation of input specifications, the screening of error types, and the formulation of tagging standards should be handed over to experts for proofreading and revision, and then handed over to machines for translation. This is the best method.

The construction of New Era People's Daily Segmented Corpus, for example, gives full play to the role of human beings. But for hundreds of millions of corpus data, professionals will be powerless. Setting up the process reasonably can reduce the need for manual proofreading

5. REFLECTIONS ON TRANSLATION TEACHING

AI-Assisted translation technology provides an important method for translation teaching and brings us inspiration. And based on the above analysis, this paper proposes the following reflections on translation teaching

5.1. The Transition from Corpus-Assisted Translation Teaching to Corpus-Driven Translation Learning.

Traditional translation teaching relies on modern translation tools to improve translation teaching. But now, as data resources become more and more open, translation teaching should also be part of translation intelligence. The translated text in computer-aided translation has also become an important source of corpus data. Under the guidance of teachers, the translation quality of professional students can improve the data quality of the translation corpus, and thus contribute more to the translation corpus. This is a typical corpus-driven learning mode, which can not only build a teaching-based corpus database, but also combine teaching with practice. We can first use bilingual parallel corpus retrieval software to retrieve typical translation cases, we analyze and think about these cases, we discuss language characteristics, translation methods and skills, bilingual conversion, translation effects, etc. Then students are asked to complete translation tasks to form bilingual corpus data and input the content into the translation corpus. Students participate in the construction of the corpus, which can better stimulate students' interest in learning. When parallel corpus is applied in translation teaching, the original text, sample translation and student translation will be presented simultaneously. Students can judge the possible problems in the translation by observing the abnormal values[13], which will contribute to the improvement of corpus database.

5.2. The Transition from Cultivating Talents who Specialize in Translation Skills to Cultivating Interdisciplinary Language Service Talents.

The development of AI will have a long-term impact on the translation industry. Take the training of translation talents for example. Translation talents not only need to master translation theories and skills, but also need to learn translation technology and knowledge about managing the language services industry. In terms of curriculum, courses on data analysis and management of translation projects should be given attention. In the teaching of MTI, translation theories and practices should be integrated.

ACKNOWLEDGMENTS

This work was financially supported by 2021 Fund for the Humanities and Social Sciences in Zaozhuang University: “Study on Chinese Cultural Words Translation from Perspective of Reception Aesthetics”(2021YB14).

REFERENCES

[1] Nature. Top Multimedia Picks of 2016 [EB/OL]. Retrieved from http: //www. Nature.com/news/top-multimedia-picks-of-2016-1.21184, 2016-12-16.

[2] Huashu Wang. (2017) “Collaborative Innovation and Normative Development of Language Services: A review of 2016 China Language Services Conference and China Translation association Annual Conference,” Chinese Translators Journal, 2017(1), 85-88.

[3] Qilu Xu.(2017) “Research on Professional Interpretation Teaching System under the Background of AI”, Technology Enhanced Foreign Language Education, 2017(10):85-88.

[4] Mingjiong Chai.(2016) “Language Service under Internet Big Data--From the AlphaGo”, East Journal of Translation, 2016 (3):4-9.

[5] Kefei Wang. An Exploration of Corpus Translation. Shanghai Jiao Tong University Press. 2012.

[6] Jiajin Xu.(2017) “The evolution of key concepts, philosophical and linguistic underpinnings of corpus research”, Foreign Language Teaching and Research, 2017, (1):51-63.

[7] Bing Xiong. (2015) “Research on translation teaching model based on an English-Chinese parallel corpus”. Foreign Language World, 2015(4): 2-10.

[8] Lihe Huang. (2015) “Corpus 4. 0: Multimodal Corpus Building and Related Research Agenda”, Journal of PLA University of Foreign Languages, 2015, 38(3):1-7.

[9] Knight D, Evans D, Carter R, Adolphs S. (2009) “Head Talk, Hand Talk and the corpus: towards a framework for multimodal, multimedia corpus development”. Corpora, 2009,4(1):1-32

[10] Hancheng Xu. (2017) “Research on Military Term Extraction Mthod Based on Corpus”. Foreign Lanuage Research, 2017(5): 43-46.

[11] Minkang Zhou. (2016) “The Enlightenment and Reference of the EU's Multilingual Inter-Translation Professional Terminology Corpus”. Chinese Translators Journal, 2016 (5):70-75.

[12] Kaibai Hu. (2016) “Research on the Characteristics of Machine Translation and Its Relationship with Human Translation”. Chinese Translators Journal. 2016(5):10-14.

[13] Kefei Wang. (2015) “The Use of Parallel Corpora in the Teaching of Translation”. Foreign Language Teaching Research, 2015(5):763-772.