Discussion on Data Features and Construction Models of Translation Corpus in the Era of Big Data

Qinqi Kang[1,a], Zhao Kang[2,b]*

[1]School of Foreign Languages, University of Electronic Science and Technology of China, Chengdu, China 611731
[2]School of Education, Research Center for Mental Health Education, China West Normal University, Nanchong, China 637009

Abstract—With the rapid development of artificial intelligence in the current era of big data, the construction of translation corpus has become a key factor in effectively achieving a highly intelligent translation. In the era of big data, the data sources and data types of translation corpus are becoming more and more diversified, which will inevitably bring about a new revolution in the construction of translation corpus. The construction of the translation corpus in the era of big data can fully rely on third-party open source data, crowd-sourcing translation, machine closed-loop, human-machine collaboration and other multiple modes to comprehensively improve the quality of translation corpus construction to better serve translation practice.

1 INTRODUCTION

The rapid development of big data technology in the current society has fully promoted machine translation to the stage of intelligent translation of neural networks. Especially with the new breakthroughs and in-depth development of artificial intelligence (hereafter referred to as AI), the AI-based machine translation has become a hot topic in translation studies in recent years. The comprehensive promotion and use of AI technology has greatly reduced machine translation errors and improved the accuracy of translation. In particular, the advent of the big data era has fully triggered a new wave of big data translation and has greatly reignited people's hope for machine translation and the construction of a new translation model that is not only closely integrated with high-tech but also adapted to the characteristics of the big data era. (Xiao Kairong, 2018) Professional translation Companies such as Google and Youdao continue to use massively statistical machine translation of parallel corpus, which completely subverts excessive machine translation or the traditional ruled translation model, thus can provide users with better services. Driven by the full power of information technology, the construction of translation corpus has become an important foundation for the current development of AI translation and will gradually become the main body of language services. (Chai Mingying, 2016) It has become an urgent call for the times development to scientifically conform to the characteristics of translation corpus in the context of current big data, fully consider the inherent relationship between translation corpus and AI translation, and actively construct new translation corpus projects that meet the requirements of big data.

2 TYPES, SOURCES AND CHARACTERISTICS OF TRANSLATION CORPUS IN THE CONTEXT OF BIG DATA

2.1 Examples of types of Translation Corpus in the Context of Big Data

Huang Lihe (2015) holds that the translation corpus in the context of big data has typical multi-modal characteristics, which is manifested in the system integration of multiple information such as audio, video and text corpus in practice. Lu Yan (2019) believes that the translation corpus in the context of big data includes types ranging from traditional terminology, text, expansion to unstructured data and knowledge graphs. The famous British scholar Mona Baker has studied and created the world's first Translational English Corpus. On this basis, he proposes three different types of translational corpus which are Parallel Corpus, Multilingual Corpus and Comparable Corpus (Zhou Xiaoling,etc,2008). In summary, the main data types of translation corpus in the context of big data are as follows:

First, professional terms translation. This includes translational professional dictionaries and other professional terminology databases, which can provide systematic professional knowledge guidance. Meanwhile, this is an important reference tool for current machine translation and can provide more accurate data information for it. Because these terms are highly

[a]kangqinqi123@126.com
[b]*Corresponding Author: kangzhao168@126.com

professional and authoritative, the construction of a professional translation term corpus should naturally become the first choice in the context of big data.

Second, text data translation. This is also an important choice for current construction of translation corpus data. In the process of translation practice, many corpus sentence pairs of machine translation rely on and mostly manifest as textual data. The various semantic information contained in these translation materials can provide fast service and application support for the retrieval and extraction of translation information. It is just because of this that machine translation around translated text data has become a key research object of current data construction of translation corpus.

Third, unstructured translation data. The diversification of data sources in the era of big data is very obvious. Various textual materials, videos and search engines from the Internet as well as translation methods such as mobile phone scanning and photographing grow rapidly. Compared with traditional text translation methods, unstructured and diversified sources of translation data will inevitably put forward more and higher requirements for the construction of translation technology.

Fourth, knowledge graph translation. That is, to use visualization technology to describe the content of translation data resources and related carriers, data mining analysis, knowledge presentation also the internal relationship between them. The rapid development of information technology will inevitably facilitate a high degree of integration of information extraction and knowledge, integrate the data that has not been connected, promote a trend of knowledge mapping in the translation corpus, and realize the effective connection of massive fragment translation data information.

2.2 Data Sources of Translation Corpus in the Context of Big Data

First of all, the data source of the translation corpus in the context of big data should and must be Internet data. The current massive amount of Internet translation data information can provide a solid foundation for refining and forming a translation corpus database. To search and extract structured data information from the disorderly and loosely structured Internet pages is an important goal of the data construction of translation corpus, for example, in practice we can make full use of the publicity homepages of different language companies to realize the crawling and automatic alignment of relevant information and documentation on the websites of multilingual companies.

Second, the translation data of companies is a very important and reliable source of information and data in the translation corpus. Many companies have built bilingual translation documents such as company profiles, product manuals and related data information, which are ready-made and available data resources. This type of data is of good translation quality, a high degree of completeness and a huge stock of assets. In practice, we can make full use of data search and data mining technology to promote those incomplete corpus data to become effective. Particular attention should be paid to strengthening cooperation with traditional text information service companies such as newspapers, publishing houses and other institutions, which can also achieve more effective translation data resources.

Third, user-generated translation data is also an important information source for translation corpus data in the Internet era. Many highly valuable translation corpus data often come from various community discussions, QQ or WeChat of users and other related information. With the help of active extraction and mining of user-generated data, it is natural to continuously enrich and improve the source of translation corpus data.

Fourth, machine-generated translation data will inevitably become one of the important information sources of translation corpus data in the era of big data. With the continuous development of information technology, especially the widespread application of AI technology, machine-generated data has become a potential data source. We can use machine translation to generate more similar sentences based on real example sentences, or use existing machine translation models to reverse-translate some monolingual corpus data to generate bilingual corpus data, which is also a very useful attempt. In fact, although this type of data has problems such as high repetition and lack of sufficient semantic information so that it is difficult to be used in large quantities. However, due to the diverse dimensions and large amount of such data, it also has its own unique existence and development space.

2.3 The Characteristics of Translation Corpus in the Context of Big Data

The construction of the translation corpus in the context of big data must fully consider the characteristics of the context of big data and the various influences it may have. Laney (2001) once did a systematic study on the era characteristics of big data and proposed that big data includes three major characteristics: volume, variety and velocity, which are the so-called 3V features. He Xiaochao (2014) proposed three V and one C characteristics of big data, including four parts: value, variability, veracity, and complexity of processing and analysis. Combining the above analysis of the types and sources of translation corpus data in the context of big data, it can be said that the translation corpus data in the context of big data presents obvious characteristics of the times, which is quite different from the traditional translation corpus data (as shown in Table 1).

TABLE 1. COMPARISON OF CHARACTERISTICS OF TRANSLATION CORPUS DATA IN THE CONTEXT OF BIG DATA

Data Sources	Data Characteristics	Data Types	Data Acquisition Methods

Internet data	The data amount is large and diverse	Term; text; unstructured data	Grab; auto-align
Enterprise data	The data acquisition is difficult	Term; text	Access the database
User-generated data	The data amount is large and with much noise	Unstructured data; text	Mine the data
Machine-generated data	The data is diverse and repeatable	Knowledge graph	Network algorithm

3 Discussion on the Construction Mode of Translation Corpus in the Context of Big Data

In the current era of big data, especially with the emergence and development of AI technology, translation technology has also developed rapidly, which puts forward newer requirements for the construction of translation corpus. Deng Zhonghua(2013) pointed out that the current scientific research of the era of big data faces four serious challenges, which are from the era of big data, the highly developed information technology, the implementation process of scientific research and the effective management and contribution of scientific data. Therefore, in the new historical period, to actively explore the construction path of the translation corpus database that meets the needs of social development has become an important subject of the development of the times.

3.1 Common Patterns of the Construction of Translation Corpus Database in the Era of Big Data

3.1.1 To Fully Rely on A Third-party Comparison Model of Open Source Data

The current existing online translation systems and translation memory systems provide a solid foundation for obtaining parallel corpus data, which can be seen from Google, Youdao, Baidu and other translation systems. During the construction process of translation corpus, many companies will fully integrate existing third-party databases and actively construct translation parallel corpus sentence pairs, which can lay a solid foundation for AI translation learning and research. In practice, the translation of the same sentence can be found in different search engines such as Google or Baidu. Based on the full combination of the translation results of different search engines, we can naturally form our own translation corpus (as shown in Figure 1). Most of these huge amounts of open source data are built on the Internet and can provide online translation services in a multi-modal combination including video and audio. For example, when translating the classic idiom “buruhuxue,yandehuzi”, different translation results will come out with the help of different translation systems. On the basis of these results, we can combine the Chinese traditional cultural characteristics and using habits to make scientific choices to translates it into "Nothing ventured, Nothing gained".

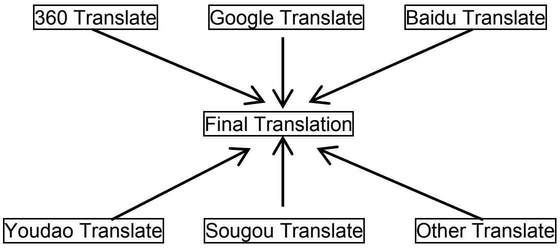

Figure 1. To Fully Rely on A Third-party Comparison Model of Open Source Data

3.1.2 Sharing Mode of Crowdsourced Translation Corpus Data

In the past few years, a new type of translation model transplanted from the American work method called “crowdsourcing” appeared in the translation industry, that is, crowdsourced translation model which means multiple translators are selected through the Internet to do translation work in the shortest time. For example, finished in the division of labor by five selected translators, the Chinese translation of "Jobs Biography" is translated and published in less than one month. Crowdsourced translation is essentially a cooperative translation, which connects individual translators and machine translation tools. It’s possible to build a corpus data sharing platform based on the organizational model of crowdsourced translation. Of course, as an objective existence in the real translation world, crowdsourced translation gets different attitude from people. In fact, the resource sharing of the translation corpus database can better solve the assets shortage of AI machine translation corpus and fully realize the high integration of corpus resources. Based on the corpus data sharing platform of crowdsourced translation, the data sharing model is shown in Figure 2.

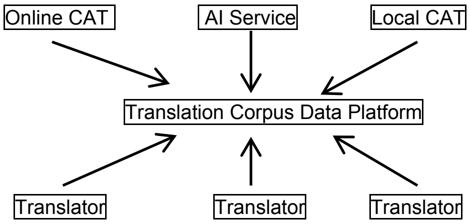

Figure 2. Sharing Mode of Crowdsourced Translation Corpus Data

3.1.3 Self-closed-loop Learning Model Based on Machine Translation

At present, the machine translation system based on neural network has the ability of self-deep learning. We can make full use of this feature to actively promote machine translation to realize cyclic learning within a certain period of time, and use model training to effectively improve both the performance and the quality of machine translation. Many practices have proven that even with the same translation corpus data content, the translation effect of the machine translation system after repeated learning can be significantly improved. This kind of self-closed-loop learning mode based on machine translation is shown in Figure 3. Of course, this also requires the review opinions of translators and user units to continuously improve the overall quality of translation.

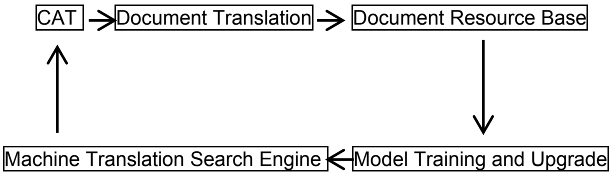

Figure 3. Self-closed-loop Learning Model Based on Machine Translation

3.1.4 Collaboration Model Based on Human-Machine Cooperation

Hutchins and Somers (1992:148) once summarized four different translation forms from a technical point of view, namely machine translation, machine-assisted translation, human-assisted machine translation and human translation. The rapid development of information technology in the era of big data has promoted the collaboration of multiple translation modes to become the inevitable trend of the development of the times. The cooperation between machine translation and human translation is an important feature of translation development in the new era (Hu Kaibao,2016). Simple machine translation in the traditional sense is naturally difficult to adapt to the real needs of the big data era. Fundamentally speaking, machine translation is also inseparable from the collaboration of human translation experts. The human-machine collaborative translation model will be an important path for the construction of translation corpus in the big data era. From the aspects of standard operation of translation corpus collection and information entry, and error correction and labeling, it is the best choice to make full use of the careful review and modification of translation experts, and then hand it over to machine translation. The mechanism of the cooperative model based on human-machine cooperation is shown in Figure 4. The human-machine cooperation model can fully realize the creative ability of human translation and the computational ability of machine translation, and promote the organic combination of the two to achieve the overall improvement of translation quality.

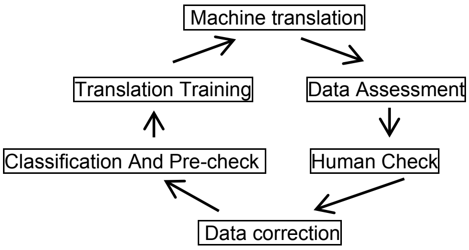

Figure 4. Collaboration Model Based on Human-Machine Cooperation

3.2 The Expected Trend of Translation Teaching and Research to Promote the Construction of Corpus Database in the Context of Big Data

After sorting out the four translation corpus database construction models mentioned above, combined with the analysis of the corpus data source types, it can be seen that different construction models of translation corpus database have their own advantages, as shown in Table 2.

TABLE 2. COMPARISON OF CHARACTERISTICS OF CONSTRUCTION MODELS OF TRANSLATION CORPUS DATABASE UNDER THE CONTEXT OF BIG DATA

Corpus Model	Applicable Data Source	Key Factors
Corpus Data Sharing	User-generated Data	Intellectual property
Third-party Open Source Data	Enterprise Data	Algorithm Design
Machine Self-closed-loop Learning	Machine-generated Data	Algorithm Design
Human-machine Collaboration	Internet Data	Process Collaboration

We should be clear that the construction of translation corpus in the era of big data is faced with practical problems such as complex data sources, diverse types and low data quality. The construction of the translation corpus database that meets the needs of the times needs to carefully combine the characteristics of various construction models of translation corpus database to maximize strengths and avoid weakness, combine organically, and actively explore new ideas for translation teaching and research in the context of big data to promote the construction of corpus databases. In fact, the construction of our country's translation corpus has been greatly developed and has been widely used in current translation teaching and research. It should be said that the development of translation technology driven by AI not only provides scientific methods for current translation teaching and research, but also puts forward more clear requirements of the times for us as follows.

First, to fully realize the learning function of translation teaching to promote the construction of the translation corpus. Traditional translation teaching often advocates relying on modern translation tools to improve teaching level. The open nature of data resources under the context of big data has promoted translation teaching to become an important part of translation intelligence and corpus. With the continuous improvement of the current students' translation level, the quality of translation data will be higher, and the contribution to the corpus will be higher. This typical corpus-driven learning model is not only conducive to the construction of teaching-type corpus databases, but also can organically integrate translation teaching and practice. Therefore, students' active participation in the construction of the corpus can better stimulate their interests in learning and meanwhile contribute to the construction of the corpus.

Second, to focus on the practical exploration from the emphasis on professional teaching of translation theory to the active construction of a vertical translation corpus database. In our country's current foreign language teaching practice, we can fully rely on the construction and development needs of relevant colleges and universities to build a vertically integrated translation corpus database to meet different requirements, which can naturally open up a new path for the organic combination of theoretical research and practical exploration of translation teaching.

Third, to strengthen and innovate the training model from translation-based professional skills talents to comprehensive language service talents. The current high-level development of big data and AI has brought an extremely far-reaching impact on the translation industry, thus promoting in-depth thinking in this industry and profound changes in the training model of translation talents. In terms of the training of translation professionals, it is necessary for them to not only master systematic professional translation theory and practical skills, but also have a more effective response to the technical drive and refined management requirements of the language service industry.

4 Conclusion

Undoubtedly, the development of information technology in the era of big data will inevitably bring about profound changes in both translation forms and translation models, thus promoting that the construction of the translation corpus database must conform to the characteristics of the times and become increasingly perfect. It should be said that the advanced development of machine translation cannot really eliminate the various boundaries between languages and human translation cannot be completely replaced by machine translation. The organic combination of multiple translation models is the basic trend of its development, which must be clearly recognized by us. Scientifically understanding and making full use of the technological advantages in the era of big data, building a more scientific and efficient translation corpus database and providing better translation services are the demands of the times for social development.

Acknowledgment

This research was financially supported by the teaching reform and research project of China West Normal University（xjjgxh2017061）.

References

1. Xiao Kai-rong. Encountering the Fourth Paradigm: Translation Studies in the Big Data Era. Foreign Language Research, vol. 2, pp. 90–95, 2018. doi:10.16263/j.cnki.23-1071/h.2018.02.014.
2. Chai Ming-jiong. Language Services of Big Data on the Internet- Speaking of Alphago. East Journal of Translation，vol. 3, pp. 4–9, 2016.
3. Huang Li-he. Corpus 4.0: Multimodal Corpus Building and Related Research Agenda. Journal of PLA University of Foreign Languages, vol. 3, pp. 1–7, 2015.
4. Lu Yan. Big Data Corpus Construction Under Artificial Intelligence Translation. Gansu Science and Technology, vol. 17, pp. 80–84, 2019.
5. Zhou Xiao-ling, Jiang Jian-song. An Analysis of the Methodology of Corpus-based Translation Studies. Journal of Xiangtan University(Philosophy and Social Sciences), vol. 4, pp. 155–158, 2008.
6. Laney, D. 3D Data Management: Controlling Data Volume, Velocity and Variety[Z]. New York: META Group, 2001.
7. He Xiao-chao. Vertical and Horizontal Big Data. Beijing: Publishing House of Electronics Industry, 2014.
8. Deng Zhong-hua, Li Zhi-fang. The Evolution of Scientific Research Paradigm:The Fourth Paradigm of Scientific Research in the Era of Big Data. Information and Documentation Services, vol. 4, pp. 20–21, 2013.
9. Hutchins,W.J.,Somers,H.L. An Introduction to Machine Translation[M]. London:Academic Press Limited, 1992.
10. Hu Kai-bao, Li Yi. Research on the Features of Machine Translation and Its Relationship with Human Translation. Chinese Translators Journal, vol. 5, pp. 10–14, 2016.