Comparative Study of Crowdsourcing Translation, Machine Translation and Computer-Aided Translation in the Information Age

Zhen'gang Yang[1,*]

[1]College of Translation Studies, Xi'an Fanyi University, Xi'an, Shaanxi, China
*Corresponding author. Email: trados@163.com

ABSTRACT

The possibilities afforded by information and computer technology in rendering the translator's work more efficiently and effectively are diverse and far-reaching. Meanwhile, with the rapid development in language service industry, translation technology itself is becoming a diversified and hybrid technology and its varied level of utilization has satisfied demands of language services industries at different levels. Crowdsourcing translation, machine translation and computer-aided translation have played critical roles in the process of translation in the information age. This paper firstly illustrates the basic concepts of crowdsourcing translation, machine translation and computer-aided translation from a comparative point of view, then demonstrates the integration and application of the above technologies in translation process and proposes that the proper understanding and application of translation technology will help to achieve the future localization paradigm that is targeted: low cost, high speed and high quality. The paper also examines the relationship between crowdsourcing translation, machine translation and computer-aided translation, and holds the opinion that they do not represent a fundamentally new form of translation, but merely a more recent stage of translation practice in the information age.

Keywords: crowdsourcing translation, machine translation, computer-aided translation

I. INTRODUCTION

As a knowledge-based activity in the information age, translation requires new strategies on how to efficiently acquire and process information. This reflects the paradigm shift of translation activities at the level of methodology and practice. It is not only limited to the field of professional translation, but will also have a certain impact on translation teaching and research [1].

Since Austermühl proposed the paradigm shift of translation activities in the information age in 1999, translation activities have been changing under the influence of technology all the time. In April 2013, the Translation Technology Landscape Report released by the Association of Users of Translation Automation (TAUS) also mentioned that the present continuous integration of economy, society and technology has brought about a paradigm shift in translation activities, from translation as a relatively expensive professional service to daily application.

The new paradigm implies a definition of translation field and language service industry that is more in line with the information age. This recognized paradigm has been guiding and influencing translation activities and related research. At present, the translated source files have completed the transformation from hard copies to digitals, software and web pages. Therefore, in order to meet the need of local customs, the process of trying to solve many issues such as language, culture, content and technology has also changed from the former translation process to Localization. Almost all professional translations are completed in a digital environment by means of computer platforms. In recent years, cloud technology, crowdsourcing, big data and Internet have all promoted the development of translation technology in different ways. The vigorous development and mutual integration of crowdsourcing, computer-aided translation and machine translation have also promoted the transformation of translation activities, thus promoting social and cultural exchanges and development. Within the context of information age, a proper understanding of the concept, relationship and application of new translation technologies is the basis for people to judge their value, and measure the role of

technology in changing the relationship between translators and texts, thus optimizing the translation process.

II. CROWDSOURCING TRANSLATION

Crowdsourcing first appeared in *Wired* magazine in the United States in June 2006. Jeff Howe, the author of this article, defined it as: "The practice of a company or organization to outsource work previously performed by employees to non-specific (and usually large) mass networks on a free and voluntary basis." In the vast majority of crowdsourcing cases, because the supply far exceeds the demand, the remuneration for such tasks is mostly free.

A. Crowdsourcing and crowdsourcing translation

Crowdsourcing does not need to rely on engines like machine translation but the support of web 2.0. With the development of web 2.0 technology, the information release on the Internet gradually changes from static information to content aggregation, and the interactive mode also changes from user to participator. Crowdsourcing involves IT, electronic consumption, tourism, photography, design and other industries. Crowdsourcing translation means that volunteers or the public voluntarily accept translation tasks published online through existing platforms, and the remuneration for the tasks is very limited if any. The process of standard crowdsourcing translation is similar to that of outsourcing translation, but there are many steps such as website homepage, login interface and registration. Registered personnel will be divided into translators, revisers, clients and administrators according to their roles, and then given the access to enter different pages. After the translation is submitted, professional interpreters will score and revise it, mark the best translation with symbols and adopt it.

Although crowdsourcing appeared in 2006, it is quite common for non-professionals to translate open source software through crowdsourcing. The earliest example is Ubuntu, a Linux operating system based on desktop applications, where users localized the system into their native languages with enthusiasm. Facebook is also an example of using crowdsourcing to translate the main platform. The entire Facebook website was translated by 4,000 dedicated French translators within 24 hours in March 2008. In the first half of 2011, American publisher Simon & Schuster released the news of the publication of "Steve Jobs: A Biography". The global publication date of the book was finally determined to be November 21, 2011, and the simplified Chinese version would also be released simultaneously. In China, it was through crowdsourcing that four translators completed 560 pages of 500,000 words in 30 days.

B. Advantages of crowdsourcing translation

Crowdsourcing translation has the advantages of multilingual support, fast speed and low cost. For low-resource languages, there is neither an automatic translation system nor professionals. Crowdsourcing translation can solve practical problems while taking into account the cost. A report released by Common Sense Advisory in February 2011 shows that crowdsourcing translation can generate multilingual translations with the help of online communities in a relatively limited time. (http://www.pr.com/press-release/300309) As many participants have access to tasks, crowdsourcing translation is faster than individual translators and team translators in traditional translation processes. In terms of cost, the initiator of the translation task needs to pay the lower maintenance cost of the crowdsourcing translation platform, while the remuneration or reward for the crowdsourcing translators will be far lower than the traditional translation cost. The advantages of crowdsourcing translation are also reflected in the wide range of target language audiences. People who use the target language as their mother tongue can enjoy translation services, such as obtaining information and using software in the mother tongue environment.

C. Challenges faced by crowdsourcing translation

The first issue of crowdsourcing translation is the quality. Kittur (2008) conducted a study on the influence of cooperation methods on quality and found that efficient cooperation skills can help improve the quality [2]. Del Giudice (2010) studied the influence of user/reader feedback and the types of praise and criticism on public trust [3]. Although crowdsourcing translation is essentially human translation, most of the participants are not certified translators, and their level varies. The establishment and improvement of the incentive mechanism for crowdsourcing translation will also be a challenge. Apart from less remuneration, hobbies, emotions, identity, sense of achievement, etc., are all fundamental factors that enable translators to participate in crowdsourcing translation. Due to the large number of participants in crowdsourcing translation, there are certain difficulties in efficient management before, during and after the translation process. Since crowdsourcing translation is done by many people through cooperation, there will be intellectual property issues such as copyright.

III. MACHINE TRANSLATION

As the initial attempt of artificial intelligence, machine translation is an automatic translation that uses computer software to translate text (or speech) from one natural language to another language to improve the quality and efficiency of translation.

A. Development of machine translation

The development of machine translation has gone through ups and downs. The first attempt of machine translation appeared in the 1930s. In 1949, Weaver, vice president of Rockefeller Foundation, published a memorandum entitled "Translation", which became the first milestone of machine translation. From 1954 to 1966, machine translation experienced a period of high expectations. From 1966 to 1976, the ALPAC report dampened the enthusiasm for the research of machine translation. Until 1970s to 1980s, machine translation was recovered and developed in Canada and Europe.

B. Classification of machine translation

According to different structures, MT can be divided into Rule-based Machine Translation (RBMT), statistical Machine Translation (SMT), Example-based Machine Translation (EBMT), Hybrid Systems, etc.

Rule-Based Machine Translation (RMBT) is the most traditional machine translation technology. Its core system is based on grammar rules, supplemented by dictionaries, including source language text processing, dictionary composition, text analysis, dictionary query translation from source language to target language, target language text generation and other technologies. The biggest advantage of RMBT lies in the management of terms and rules, including the establishment of grammar rules that can effectively speed up the construction of new projects. Disadvantages are that the establishment of grammar rules is very time-consuming, and the target translations generated are mostly literal translations without enough consideration for the context. The rule-based machine translation system is difficult to continue to improve after reaching a certain level.

The Statistical-based Machine Translation (SMT) is to make statistical analysis of the two languages by sorting out the database (bilingual corpus) to construct a reasonable statistical translation model, based on which the model parameters will be estimated and defined, and the parameter estimation algorithm is designed, thus realizing automatic translation between languages. SMT has the advantages that the generated target text is more readable than that of RMBT with shorter setting and adjustment time and enhanced sensitivity to the context. The translation can reflect the style of the statistical corpus. However, it requires a large amount of bilingual corpus and takes a long time to process the corpus. Most of the languages supported by Google Translation adopt statistical machine translation methods, making machine translation receive unprecedented attention.

Example-based Machine Translation (EMBT) was proposed by Makoto Nagao, Japan. The basic idea is to find out the most similar translation examples (usually sentences) to be translated from the collected bilingual corpus, and then generate the translation through a series of operations such as string replacement, string deletion and string addition to realize the translation process. EMBT has the advantages of easy system maintenance and high quality. However, there are high requirements for bilingual corpus alignment technology and quality, as well as efficient retrieval and matching technology [4].

Hybrid Machine Translation (Hybrid Systems), also known as a multi-engine machine translation system, integrates various machine translation methods such as dictionary, vocabulary conversion, rule-based and instance-based machine translation into one environment. In the translation process, multiple translation engines process the input sentences simultaneously or separately, and generate the final target text through two different working modes of pre-translation judgment and post-translation judgment [5].

At present, online machine translation has been widely used. Online closed source machine translation systems, such as SYSTRAN, PROMT, can also be used on local computers, and some also provide mobile phone applications, targeting potential buyers of closed source translation software. Online machine translation services are mainly aimed at the general public. For example, Yahoo Babelfish is based on SYSTRAN and is not the official closed source version of SYSTRAN. Therefore, online translation services are subject to some restrictions. Google translation tools can deal with translation in more than 50 languages; Microsoft Bing Translation, similar to Google's online translation, also provides users with scoring and post-translation editing functions through WYSIWYG editing tools.

C. Development prospect

While understanding machine translation, we should pay attention to the common misunderstanding of machine translation: poor quality, less creativity, negligence of style, too complicated technology and so on.

According to research by Melby and the viewpoint of Dimitra Anastasiou, it is believed that the traditional RBMT system will not receive much attention in the next few years. Corpus-Based SMT will get better development and attention [6], [7]. The future machine translation system will take hybrid machine translation combined with RBMT, SMT or EBMT as the development mainstream. The combination of open source technology, data sharing, and machine translation and human translation will be the new development trend of machine translation. Machine translation and speech recognition technology will be further combined, and real-time speech machine translation system will be further developed and applied in the commercial field.

IV. COMPUTER-AIDED TRANSLATION

The technology of computer-aided translation (CAT) originates from machine translation. It is a combination of human translation and machine translation with the help of computer software and hardware technology under the background of information age.

A. CAT in broad and narrow sense

The concept of computer-aided translation has been discussed by scholars both at home and abroad. Lynne (2002) mentioned that, in a broad sense, computer-aided translation technology includes all computer tools that can help translators to complete the task, including word processing, grammar checking, e-mail and network resources, and the above technologies are becoming general knowledge nowadays [8]. Yu Jingsong and Wang Huashu (2010) further enriched the broad concept of computer-aided translation technology, and proposed to include translatable resources extraction, task analysis, terminology extraction, fragment reuse, terminology identification in translation, format conversion, language asset management, etc. [9]. In a narrow sense, computer-aided translation technology mainly refers to the use of translation memory technology and terminology management technology to improve translation quality and efficiency.

B. Translation memory and terminology management technology

As the core technology of computer-aided translation in a narrow sense, translation memory refers to a database capable of memorizing translated sentence segments, which will automatically identify repeated content, display matching rate to improve translation efficiency. Terminology management technology ensures consistency in the use of translated terms and improves translation quality through terminology database, a database containing indexes, terms and related information. There are many mainstream computer-aided translation software, such as SDL Trados, Déjà vu, MemoQ, etc.

C. Development trend of computer aided translation

Computer-Aided translation has gone through the development process from the original stand-alone software application to server application and then to cloud technology. Computer-aided translation system has further integrated with content management system. For example, SDL World Server, an industry-leading translation management system, combines mainstream content management systems Drupal and Sharepoint to realize efficient management of content, editing and publishing workflows. Mobile device translation tools have also become a new trend in the development of computer-aided translation technology.

V. CROWDSOURCING TRANSLATION, MACHINE TRANSLATION AND COMPUTER AIDED TRANSLATION: HOW TO BALANCE PRICE, SPEED AND QUALITY

A. Comparison of the three

The biggest difference between crowdsourcing translation and machine translation is that crowdsourcing translation is human translation in which multiple translators cooperate, while machine translation is automatic translation without human intervention. Crowdsourcing translation requires real-time human participation or intervention, while machine translation requires human intervention during text input and post-editing. Crowdsourcing translation involves more translators, thus requiring translation process management. Lynne also proposed that the main difference between machine translation and computer-aided translation lies in who is the main body to complete the translation task. In the computer-aided translation environment, human is the main body, and various computer tools will be used to help complete the translation tasks and improve the efficiency in the translation process. The development trend of machine translation is to try to replace human translation, while computer-aided translation is to help improve the efficiency of it.

B. Future localization paradigm: balance among price, speed and quality

Dimitra pointed out that the future localization paradigm is the combination of automatic translation and human translation, taking into account of low price, high speed and high quality ("Fig. 1"):

Fig. 1. The future localization paradigm.

In terms of speed, machine translation is the fastest; Crowdsourcing translation is a collaborative work of multiple translators, ranking the second. Computer-aided translation is faster than traditional human translation due to the use of new technologies.

In terms of cost, machine translation prevails. Thicke proposed that machine translation broke the

compromise between price, speed and quality, leaving us with only two choices [10]. Machine translation can provide services that meet the above three conditions in some cases. In fact, for most people, the use of machine translation still requires choosing two from price, speed and quality. Crowdsourcing translation platform is a relatively low-cost solution, which is cheaper than machine translation. Crowdsourcing projects require budget in the initial stage, and there is almost no cost in the later stage. Computer-aided translation can significantly reduce costs when there are duplicate content and similar items.

In terms of quality, because human translator can easily handle complex syntax and semantics, the quality is generally higher than that of machine translation and its application field is wider. Crowdsourcing translation is essentially human translation, and its quality is inferior to that of traditional human translation due to the influence of more translators and working modes. The quality of machine translation depends on the selection and optimization of translation engines. The scope of application is limited and the quality of translation is unstable. The essence of computer-aided translation is human translation and the quality is high with translation memory and terminology technology

C. Quality requirement

The quality requirement depends on: what kind of translation quality does a project need, understandable or publishable? What kind of translation process are required, human translation, human translation plus machine translation? Accuracy, consistent style, consistent terminology and correct spelling are all high-quality components. Of course, different styles have different quality requirements. For software translation, quality requirements should include accuracy, understandability and timely release. For product catalogues, the accuracy of product-related terms becomes the primary factor of quality standards; for the instructions, high quality means that by following the translated texts, the customer can successfully solve the problems.

D. Giving full play to the advantages of the three

The problems of crowdsourcing translation lie in the large number of translators, their level of competence, the control of the translation process and the consistency of terminology. To solve these problems, it is necessary to strengthen the selection of translators at the beginning of the project to ensure quality personnel. Using the translation management system and providing feedback during the translation process are effective to strengthen the consistency of terminology. Proofreading and scoring by professionals is also an effective way of quality assurance. Vamshi Ambati studies and proposes a collaborative work flow for

crowdsourcing translation. The three stages include contextual vocabulary translation, sentence translation and sentence combination. The quality of the translation in the previous stage can be strengthened in the subsequent stages [11].

In order to give full play to the advantages of machine translation, attention should be paid to selecting the appropriate translation engine, optimizing the translation engine and editing before and after translation. Machine translation can achieve high quality in specific fields, such as software, operation manuals, product catalogues and other technical documents. This advantage should be fully utilized. Selecting a suitable machine translation engine before pre-translation processing, optimization or training of machine translation is also crucial for improving translation quality. For example, for statistical machine translation systems, training requires not only a large number of high-quality bilingual texts, but also vocabulary and single-language versions. During post-editing, make clear the quality standard requirements, edit appropriately according to the process, and make maximum use of machine translation results.

The advantage of computer-aided translation lies in its ability to combine the latest achievements of machine translation and cloud technology through the Internet, make full use of translation memory and terminology management technology, and significantly improve the efficiency and quality of human translation. Computer-aided translation will play a vital role in the establishment of a free translator's working platform, language asset management, terminology management, etc.

VI. CONCLUSION

In the information age, the rapid development of translation technology and its mutual integration have had a great impact on translators, making them more and more dependent on the translation technology in the information age. Those who cannot adapt to the new paradigm shift will gradually lose their competitiveness. Crowdsourcing translation, machine translation and computer-aided translation have further integrated with translation technology, attracting the interest of many researchers at home and abroad. This paper summarizes the concepts of three translation techniques. By analyzing, comparing and taking the modern localization paradigm of "taking into account price, speed and quality" as the goal, this paper discusses how to combine the advantages of the three and optimize the translation process.

Crowdsourcing translation, machine translation and computer-aided translation are playing an important role in the modern language service industry, while they are not the new forms of translation, but the new stage of translation practice in the information age.

However, at the same time, translation technology brings more resources to translators in the context of the information age, and is changing the relationship between translators and texts. On the one hand, information technology and computer tools make the relationship between translators and texts more flexible, which is different from the traditional paper document translation assisted by printed dictionaries and materials. On the other hand, the Internet has changed the way of communication, providing interpreters with a way to communicate freely with the outside world without geographical restrictions, enabling real-time communication with each other, which is unprecedented.

Although these changes have failed to create new positions in the translation industry and transform the role of translators into editors, nor pose a threat to translation practitioners, we should realize that the changes brought about by technology to the translation industry are objective and will continue to bring about impacts. Therefore, the most important thing is to accept, recognize and learn new technologies, combine and exert their unlimited potential, improve the quality and efficiency of translation, and optimize the translation process.

References

[1] Austermühl, F. (1999). "Between Babel and bytes. The discipline of translation in the information age". AREAS Annual Report on English and American Studies, 16, 439–450. Trier: WVT Wissenschaftlicher Verlag Trier. (http://www.hit.uib.no/AcoHum/abs/Austermuehl.htm)

[2] Kittur A and Kraut RE. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work, New York, 2008, pp. 37–46.

[3] Del Giudice K. Crowdsourcing credibility: the impact of audience feedback on Web page credibility. Proceedings of the American Society for Information Science and Technology 2010; 47(1): 1–9.

[4] Wang Xiaojie. Fundamentals of Natural Language Processing Technology. Beijing: Beijing University of Posts and Telecommunications Press, 2002.

[5] Zhang Chunxiang. Acquisition of Translation Knowledge Based on Phrase Evaluation. Harbin: Harbin University of Technology Press, 2012.

[6] Melby AK. MT+TM+QA: the future is ours. Traduma´tica 4 2006, http://www.fti.uab.es/tradumatica/revista/num4/articles/04/04art.htm.

[7] Dimitra Anastasiou, Comparison of crowdsourcing translation with Machine Translation, Journal of Information Science, 37 (6) 2011, pp. 637–659.

[8] Lynne Bowker, 2002. Computer-Aided Translation Technology. pp. 4-6.

[9] Yu Jingsong and Wang Huashu. Discussion on the Teaching of Computer Aided Translation for MTI [J]. Chinese Translation Journals. 2010(03), P40-44.

[10] Thicke L. Optimized MT for higher translation quality. Multilingual October/November 2009: 9-11.

[11] Vamshi Ambati, CSCW'12 Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. pp 1191-1194.