Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review

Kai Jiang
College of Foreign Languages
Huazhong Agricultural University
Wuhan, China
jiangkai@mail.hzau.edu.cn

Xi Lu
Department of Common Required Courses
Hubei Institute of Fine Arts
Wuhan, China

2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI) | 978-1-7281-7738-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/IICSPI51290.2020.9332458

***Abstract*—As an essential part of artificial intelligence technology, natural language processing is rooted in multiple disciplines such as linguistics, computer science, and mathematics. The rapid advancements in natural language processing provides strong support for machine translation research. This paper first introduces the key concepts and main content of natural language processing, and briefly reviews the history and progress of NLP research at home and abroad. Then, the paper summarizes the three stages of machine translation and its research status. In history, the progress curve of natural language processing almost accords with that of machine translation, and the two complement each other. Based on this, the paper analyzes the applications of natural language processing in machine translation, and points out the challenges and trends in the field of natural language processing. Finally, the author discusses the relationship between machine translation and human translation in the age of artificial intelligence, and visualizes the future prospect of machine translation.**

Keywords—natural language processing, machine translation, artificial intelligence, translation technology, machine learning

I. Introduction

Natural language processing (NLP) is an important branch in the field of computer science and artificial intelligence. Natural language processing research comprises a wide range of theories and methods that aim to achieve effective and efficient communication between human and machine through natural language. NLP is the joint field of computer science, artificial intelligence, and linguistics that focuses on the interaction between machine and human. Language information processing, or machine translation (MT), is the earliest application of computer technology in non-numerical aspect. As the progress of artificial intelligence technology, natural language processing provides strong support for machine translation research. Under the current trend of artificial intelligence, machine translation theory and technology have received increasing public attention [1]. It is worth noting that there still exist biased opinions about the impact of natural language processing on machine translation and fear that machine translation will replace human translation [2].

In view of these, the paper first reviews and analyzes the key concepts and progress of natural language processing, and summarizes the developmental stage and research status of machine translation. Based on this, the author discusses the application of natural language processing in machine translation. Then, the paper explores the relationship between machine translation and human translation, and looks into the future trend of natural language processing and machine translation.

II. The Key Concepts and Main Content of Natural Language Processing

A. The Key Concepts of NLP

Natural Language Processing (NLP) is a subject that studies the language communication issues between human and machine. In 1998, Bill Manaris gave its definition in *Advanced in Computers*, "Natural language processing is defined as a discipline that studies language issues in human-to-human and human-to-machine communication" [12]. Its tasks can be divided into two sections—natural language understanding (NLU) and natural language generation (NLG).

B. The Main Content of NLP

Based on the viewpoint of linguists, language comprises the following levels: phonetics, lexis, grammar, semantics, discourse, and pragmatics. The applications of natural language processing on the above levels can be further subdivided into these sections: machine translation, sound recognition, sound synthesis, automatic information retrieval, term database, optical character recognition, man-machine dialogue, etc [2].

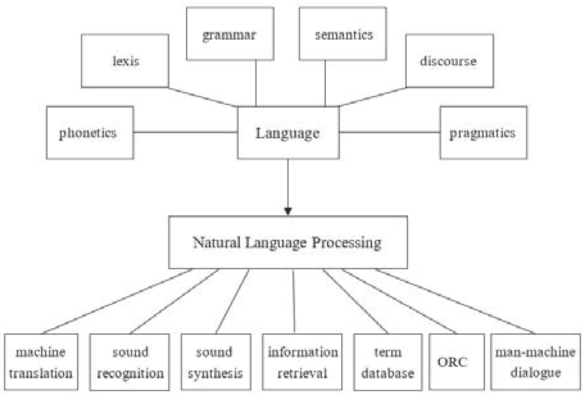

Fig. 1. The components of language and applications of NLP

At present, natural language processing has gained abundant accumulation in terms of theoretical basis, language resources, and key technologies [4]. The above-mentioned applications have also been greatly progressed. For instance, sound recognition is used in speech translation and interrogation systems at unmanned service centers at railway stations or airports. Optical character recognition technology, which can be applied to scanning software, is adopted to

978-1-7281-7738-0/20/$31.00 ©2020 IEEE

recognize printed fonts and even handwritten scripts, and ultimately generate electronic documents. Furthermore, natural language processing is extensively utilized in many aspects of translation tools.

III. The history and Development of Natural Language Processing

A. The Initial Stage (1940s-1950s)

The American mathematician Warren Weaver first proposed the use of computer in language translation. Weaver regarded translation as a process of decoding, and attempted to convert word by word through intermediate language. Yet since translating natural language is not limited to translating words, emphasis should also be laid on issues such as grammatical structure and semantic analysis. Therefore, the use of "decode" method could cause fragmented sentences, semantic contradictions, etc. This prompts researchers to improve the quality of machine translation. The division and reorganization of grammatical structure and semantic analysis have subsequently become the top issues in machine translation. During the 1950s and 1960s, there were two major trends in the development of natural language processing. Based on the different methodologies, NLP research was divided into two schools: the symbolists and the frequentists [3].

Symbolists insist on a complete and comprehensive analysis of natural language processing, and its process is highly accurate and integrated. The first parsing system came from the Transformation and Discourse Analysis Project conducted by Zellig Harris, a representative of descriptive linguistics at the University of Pennsylvania. His theories and methods to some extent reflect the features of the semiotic view on natural language processing.

Frequentists are mostly professional researchers in statistics. They search, organize, and analyze machine translation data, use probability statistics to speculate on the results of natural language processing, and widely applies the classic method of calculating hypothetical probabilities—Bayesian Analysis. In 1959, Bledsoe and Browning devised a Bayesian system for text recognition, striving to achieve the optimized recognition and computation of natural language.

B. The Developing Stage (1960s-1970s)

In the 1960s, the French mathematician Bernard Vauquois at the Centre d'Étude pour la Traduction Automatique of the Institut Mathematique Appliquée de Grenoble divided machine translation into three major steps: the analysis of morphology and syntax in the source text, the conversion of lexis and grammatical structure from the source text to the target text, and the generation of lexis and syntax in the target text. It constituted a complete set of machine translation procedures and was applied to the translation between French and Russian.

During the same period, the semantic analysis and selection of input text was not neglected. In 1974, British artificial intelligence expert Y.A. Wilkes proposed "Preference Semantics", emphasizing that the computer translation of natural language should always put semantic issues in the first place. Based on this, Wilkes designed an English-French translation system, which has excellent processing performance in semantics and the translated text is highly readable. Furthermore, the application of logic methods also attained some achievements in natural language processing. In the 1970s, Alain Colmerauer at the Aix-Marseille Université developed the Prolog language and its system (Prolog is the basic language and logic system for natural language processing and system programming), and designed the Q system and transformational grammar by using logic methods. These were then applied to machine translation. In 1972, Terry Winograd devised the SHRDLU system at the Massachusetts Institute of Technology, which integrated language analysis with knowledge reasoning, and was regarded as a major step in the progress of natural language understanding. In 1970, William A. Woods put forth Augmented Transition Network (ATN), and in 1972 he built the LUNAR system at BBN in the United States. At present, ATN has become a widely adopted method in natural language understanding research.

Nonetheless, the development of natural language machine translation once entered a period of stagnation during the 1970s and 1980s. At that time, the amount of information in computer corpus was limited, and the theory and technology of natural language processing were immature. The European Community, the United States, and the former Soviet Union all had huge capital investment, but it did not bring about substantial innovation or breakthrough in natural language processing.

C. The Booming Stage (1990s-present)

In the 1990s, natural language processing gradually entered a period of prosperity. The Machine Translation Summit IV held in Kobe, Japan in 1993 marked a new era of natural language processing. During this period, the field of natural language processing has two distinct features: large scale and usability.

Large scale means that computer has higher requirements for natural language processing. As for the input of textual information, the computer must have the capacity to process a large amount of texts, rather than a single or fragmental sentence. This requires the construction of large-scale corpora to increase the amount of texts processed by the computer system. Usability is to improve the availability of natural language processed data, and ultimately satisfy the demands for automatic retrieval and extraction. Therefore, a large database of real-life texts can serve as a "dictionary" for natural language processing. It can be seen that large scale and usability complement each other. In fact, the reason why natural language processing can pass through the "stagnation" period and revive is because of the collaboration between statistics and computer science that empowers machine to discover "features" from large amounts of data and learn from them.

Since the mid-1980s, China started large-scale and systematic research on natural language processing, and has attained achievements in certain research fields, including the construction of data resources (e.g. corpus and knowledge base), supporting technologies (e.g. word segmentation and syntactic analysis), and applied technologies (e.g. information retrieval and machine translation).

IV. The History and Development of Machine Translation

Machine translation is the use of computer to realize automatic translation between different languages. Generally, the original language is called the source text, and the translated language is called the target text. Machine translation is the conversion process from the source text to the target text. Machine translation began in the 1920s, thereby having a history of nearly 100 years. In the early 1930s, French engineer George Artsouni proposed the idea of using machines for translation. In 1933, George Artsouni and Petr Smirnov-Troyanskii applied for a patent on machine translation, but at that time machine translation simply used mechanical device to do translation at lexical level [13]. In 1947, Warren Weaver proposed the utilization of computer in translation. Afterwards, as the rapid advancements in computer technology, academic scholars began to study how to utilize computer in translation. In general, the development of machine translation consists of three stages: rule-based machine translation, statistical machine translation, and neural network translation [5].

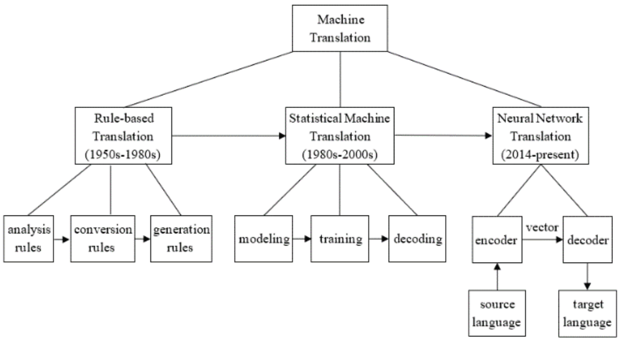

Fig. 2. The history and development of machine translation

A. Rule-based Machine Translation (1950s-1980s)

Rule-based machine translation relies on the analysis rules of the source language, the conversion rules between the source and target languages, and the generation rules of the target language. These rules involve different linguistic levels such as lexis, grammar, and semantics. Since the 1950s, rule-based machine translation has become the primary form of machine translation. The core of rule-based machine translation lies in the description and construction of these rule systems. The quality of machine translation directly depends on the scope, depth and applicability of the rules. Yet it is not easy to formulate rules that can describe all language phenomena, and often there exist overlaps or contradictions between the rules. In this sense, rule-based machine translation has its innate limitations, and its translation quality is barely satisfactory.

B. Statistical Machine Translation (1980s-2000s)

Statistical machine translation refers to the translation based on statistical models that are constructed by a large amount of parallel corpus analysis. Its theoretical basis is that any sentence in the target language may be the translation of a sentence in the source language [15]. The task of machine translation is to convert the source text into the target text through model construction. Statistical machine translation mainly involves modeling, training, and decoding. In 1949, Warren Weaver proposed the use of statistical machine translation method. Statistical machine translation was initially based on words, and then transferred to phrases, and gradually incorporated syntactics, thereby greatly improving the quality of translation. However, since statistical machine translation needs large-scale bilingual corpus, machine translation in the general field is rarely based on statistical methods.

C. Neural Network Translation (2014-present)

Neural network translation has progressed rapidly in recent years. It mainly consists of two modules: the encoder and the decoder. The encoder converts the source language into a high-dimensional vector after a series of neural network transformation. The decoder is responsible for decoding (translating) this high-dimensional vector into the target language [14]. As the advancements in deep learning technology, neural network translation began to emerge since 2014. In 2015, Baidu released the world's first Internet neural network translation system. In 2016, Google announced Google Neural Machine Translation (GNMT). In 2016, Microsoft launched Microsoft Translator based on neural machine translation (NMT) technology. In merely five or six years, the quality of neural network translation in most languages has surpassed that of statistical machine translation.

V. The Applications of Natural Language Processing in Machine Translation

In recent years, natural language processing technology has achieved rapid progress, and a series of machine translation systems have been devised and widely used. In 2006, Google began to develop its machine translation system, and finally brought Google Translate to the market. In 2011, Baidu launched its machine translation system that can support 27 languages. These machine translation systems are extensively utilized in our daily work and life, and the quality of translation has been enhanced significantly [2].

A. Machine Learning

A distinct feature of natural language processing is that it now increasingly adopts machine learning methods to acquire language knowledge [6]. Machine learning is a discipline to study how to exploit experience to enhance the performance of the system through computational means [8]. Specifically, the computer derives the "model" algorithm from a large amount of data, and then provides the empirical data to the computer, from which it generates a new model. Eventually, when new data appear, the computer can make corresponding judgments based on the generated model.

B. Machine Translation

Machine translation is a system that utilizes the computer to translate, that is, a computer program which is designed to translate text from one language (source language) to another language (target language) without the help of human [9]. Machine translation constitutes one of the most important applications of natural language processing—a field which combines the elements of information technology with linguistics. It involves many classic natural language processing issues during the operation, such as data mining and cleansing, word segmentation, part-of-speech tagging, and syntactic analysis. In addition, machine translation is also closely related to the application of machine learning algorithms.

Machine translation is broadly divided into rule-based machine translation and corpus-based machine translation. According to the different modeling methods, corpus-based machine translation is further divided into example-based machine translation, statistical machine translation and neural machine translation. In machine translation, corpus is referred to as data, which means that machine translation requires a large amount of corpus to train the model. Different types of corpus are utilized to train different models. For instance, target language corpus is used to train language generation models (to improve sentence fluency), and parallel corpus is used to train translation models (to acquire translation knowledge).

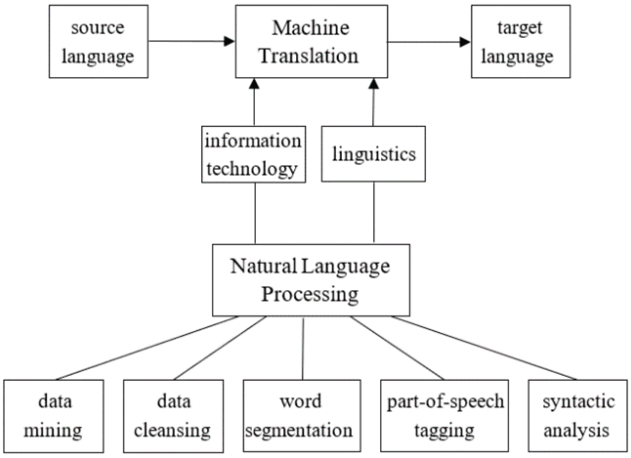

Fig. 3. The application of NLP in machine translation

As mentioned above, machine learning algorithm is utilized both in probability modelling and model training. The model building in machine translation also needs parameters to further optimize the algorithms. It can be seen that machine translation is an important research branch of natural language processing. Through the analysis, it is obvious that natural language processing technology plays a pivotal role in machine translation.

VI. Challenges and Trends of Natural Language Processing

A. Challenges Faced by Natural Language Processing

At present, natural language processing faces four major challenges: 1) widespread uncertainty: covering phonetic, lexical, syntactic, semantic, and pragmatic levels; 2) unpredictability of language phenomena: new words, terminology, semantics are emerging rapidly; 3) data insufficiency: a set of limited linguistic data cannot comprise all the languages in real life; 4) complexity of language generation: the ambiguity and intricacy of semantic knowledge cannot be described by simple mathematical models, and semantic computation demands nonlinear computation methods with complex parameters [10].

By adopting mathematical tools of modern statistics to construct complex statistical models, we can incorporate intricate linguistic knowledge (e.g. syntactics.) to effectively solve remote factors. For statistical model to achieve greater success, it depends on the breakthrough of linguistic theory and the support of massive language resources. The current research on natural language processing abroad shows the following distinct features: 1) the rationalist approach based on syntactic-semantic rules is questioned, and the processing of large-scale real-life texts has become the main strategic goal of natural language processing; 2) automatic machine learning methods are commonly adopted to acquire language knowledge; 3) the statistical mathematics is valued; 4) a strong lexicalism trend has emerged, emphasizing the function of lexis in natural language processing.

B. The Development Trends of Natural Language Processing

In recent years, natural language processing enters a stage of rapid progress. Lexicons, semantic and syntactic dictionaries, corpora and other digital resources become increasingly abundant. Related technologies such as word segmentation, part-of-speech tagging, and syntactic analysis are fast advancing. The emergence of new theories, new methods, and new models has brought prosperity to natural language processing research [2].

First, the expansion of database. Large-scale bilingual data improves the translation quality, and the model built on it can learn a wide range of translation knowledge. As the advancements in information technology, large amounts of data are collected. These data can be utilized as training corpus, thereby bringing further possibilities to machine translation.

Second, the upgrade of algorithm. The optimization of algorithms has played a pivotal role in the development of machine translation. With the rapid progress of artificial intelligence, deep learning has achieved impressive results in machine translation. The neural machine translation based on deep learning method has considerably improved the translation accuracy and fluency. As the continual advancements in computing power and algorithm, neural machine translation will become highly advanced and efficient.

Third, the advance on computing capacity. The integration of high-performance computing capacity with machine translation technology further enhances translation quality. For instance, the acceleration of computation shortens the time-delay of speech translation. And customized computation can accomplish diverse computing tasks. Applying these computing tasks to machine translation products will facilitate the use of translation technology in daily life [2].

VII. The Future Trend of Machine Translation: Machine Translation Post-Editing

As the progress of big data and artificial intelligence technology, machine translation based on neural network becomes a major focus in the society. Internet companies have launched machine translation services one after another, and cloud translation platforms have also sprung up. At present, the translation industry is developing at an unprecedented speed. Technological innovation and explosive demand for translation services have prompted machine translation post-editing (MTPE) to become the main business mode of language service providers [11].

Machine translation post-editing means that professionally trained translators modify the original translation done by machine according to the client's requirements, so as to improve its quality and applicability. The MTPE model achieves the balance between translation quality and efficiency, and meets the needs of the fast-growing language service market. *The Language Services Market: 2016* published by CSA Research in the United States shows that MTPE has become the mainstream choice for companies to provide language services to clients around the world.

Machine translation is an automated cross-language conversion process, in which the translating speed and scale are far beyond the reach of human translator. However, machine translation is a simplistic mechanical simulation. Its creativity in translation is far lower than that of human translator, and it cannot convey interpersonal meaning or pragmatic connotation. After analyzing the principle of machine translation and its similarities and differences with human translation, it can be seen that the future relationship between machine translation and human translation is not a contradiction or zero-sum competition, rather the two complement and reinforce each other [5].

VIII. Conclusion

Natural language processing is rooted in multi-disciplines such as linguistics, computer science, and mathematics, and it has now become a major research field of artificial intelligence. The rapid advancements in natural language processing provides strong support for machine translation research. Natural language processing, a core component of artificial intelligence, began with machine translation. Historically, the development of natural language processing almost accords with that of machine translation, and the two complement each other [7]. Natural language processing has been widely utilized in machine translation, and new breakthroughs have been achieved. This not only opens up a broad scope for machine translation research, but also adds vitality to its progress. This paper introduces the history and development of natural language processing and machine translation, analyzes the applications of natural language processing in machine translation, and discusses the challenges faced by natural language processing and the future trend of machine translation. In the era of knowledge-driven economy, with the continual in-depth research on the cognitive mechanism of human brain and the advancements in natural language processing, the quality and efficiency of machine translation would be substantially increased.

References

[1] M. Li, S.J. Liu, D.D. Zhang, M. Zhou, Machine Translation. Beijing: Higher Education Press, 2018.

[2] Y. Wang, “Natural language processing and applications in machine learning,” Modern Chinese, vol. 5, pp.187-191, 2019.

[3] Y.F. Song, “The development history and current situation of natural language processing,” China High-Tech, vol. 3, pp.64-66, 2019.

[4] M. Wang, S.W. Yu, X.F. Zhu, “Natural language processing and its applications in education,” Mathematics in Practice and Theory, vol. 40, issue 20, pp.151-156, 2015.

[5] K.B. Hu, Y. Li, “The features of machine translation and its relationship with human translation,” Chinese Translators Journal, vol. 37, issue 5, pp.10-14, 2016.

[6] Z.W. Feng, “Parallel development of machine translation and artificial intelligence,” Journal of Foreign Languages, vol. 41, issue 6, pp.35-48, 2018.

[7] Z.W. Feng, “Computational linguistics: its past and present,” Journal of Foreign Languages, vol. 34, issue 1, pp.9-17, 2011.

[8] Z.H. Zhou, Machine Learning. Beijing: Tsinghua University Press, 2016.

[9] W.J. Hutchins, Machine Translation: Past, Present, Future. Chichester: Ellis Horwood Limited, 1986.

[10] Q.P. Jiang, “Challenges and future of natural language processing”, China Computer & Communication, vol. 14, pp. 219-221, 2013.

[11] B. Manaris, “Natural language processing: a human-computer interaction perspective,” Advances in Computers, vol.47, pp.1-66, 1998.

[12] Y. Bar-Hillel, “The present status of automatic translation of languages,” Advances in Computers, vol.1, pp.91-163, 1960.

[13] I. Sutskever, V. Oriol, V.L. Quoc, “Sequence to sequence learning with neural networks,” Advances in Neural Information Processing Systems, vol. 4, pp. 3104-3112, 2014.

[14] P.E. Brown, J.D. Vincent, A.D. Stephen, L.M. Robert, “The mathematics of statistical machine translation: parameter estimation,” Computational Linguistics, vol. 19, issue 2, pp.263-311, 1993.