The Translation of Historical Documents and the Study of Korean History Using Artificial Intelligence

Geunhye Hong*

Introduction

At the 2016 World Economic Forum, Klaus Schwab presented the "fourth industrial revolution" as a topic for discussion. According to him, human life is expected to face radical and fundamental changes that may even be called a "revolution" due to new technology represented by artificial intelligence (AI), robotics, and the Internet of Things (IoT). His argument was that the fourth industrial revolution had already begun and we must prepare for devastating innovation and change.[1]

Shortly after Schwab's claim, national interest in Korea in the fourth industrial revolution was heightened by the March 2016 showdown between Lee Sedol and AlphaGo. Then, the fourth industrial revolution was raised as a major issue in the wake of Korea's 2017 presidential election. Later, the government responded to the fourth industrial revolution on a policy basis, the press started publishing articles on it, and a series of books have come out on the topic. Meanwhile, some people expressed critical views that the fourth industrial revolution was nothing more than a buzzword without substance.[2]

*   Ph.D. candidate in Korean History at Korea University

1 Kluas Schwab, *The Fourth Industrial Revolution* (London: Penguin Press, 2017).

2   Kim Soyoung et al, *The Ghost of the Fourth Industrial Revolution* (Seoul: Human-

A variety of discussions related to the fourth industrial revolution have been under way, including raising the fundamental question of whether it is a valid concept. However, what is clear is that preparedness for the future of mankind is necessary for its survival. Everyone now feels that human life is changing at a rapid pace. Although the current discussion is limited to the concept of the fourth industrial revolution, the sense of crisis given by the phrase "fourth industrial revolution" serves as an important opportunity for mankind to prepare for the future and reflect on the present. If we recall that Schwab's underlying intention of mentioning the fourth industrial revolution was to cope with the rapid changes in the world, it must be a significant agenda point at present.

The concern about the future of mankind is not just a concern for natural scientists but also for scholars in the humanities. Therefore, historical studies cannot be excluded from the changes that the fourth industrial revolution will bring. Even from a historical studies point of view, it has embraced ideological changes and built its own academic field based on changes and trends over time. So, it is necessary to reflect on how the coming fourth industrial revolution will impact the study of history, and the discipline must prepare for the role that historical studies and historians can or should play within that trend. History is the study of the past, but the study of the past leads to reflection on human life in the present and becomes the cornerstone of the future. Therefore, historical studies should predict and prepare for the coming era based on an awareness of

ist, 2017). In their book *The Ghost of the Fourth Industrial Revolution*, Kim So-young and co-authors criticized the fourth industrial revolution, saying it was only consumed as political rhetoric that emphasized crisis rather than a basic interest. Hong Ki-bin also criticized the fourth industrial revolution theory, drawing attention to certain information and communication technologies instead of the general science and technology in society, and expressed the view of technological determinism, according to which if such technologies are developed, industries will develop and society will change. They argue that the role of ICT and science should be adjusted, focusing on efforts that reasonably change society rather than just thinking of science as a tool for technological development.

the present, which is more intense than in any other discipline. Thus, within the trend of the fourth industrial revolution, historical studies need to explore that revolution and contemplate the future.

The development of AI in the fourth industrial revolution is a key technology and feature. AI is expected to be used in various fields and to change human lives. Among such changes, a technology that stands out for its remarkable development and usefulness is translation using AI. Since history is a study that uses historical records, these records have to be translated first in order for them to be useful to historians not familiar with the language they are written in. So, it is necessary to think about how translation using AI will be carried out, how it can be used in history, and how historians should prepare. Therefore, this paper will look at the necessity and current status of historical document translation from Chinese to Korean, using AI, and make suggestions about what roles historians can play.

The Advantages of Translating Historical Documents Using Artificial Intelligence

Although the concept of the fourth industrial revolution is rather vague, in general it is understood that the development of the science and technology of AI is one of its main concepts. There is a lot of controversy among scholars about how intelligent AI can become and how fast it will develop. The majority of scientists predict that AI will have a superhuman intelligence.[3] They argue that in the era of the fourth industrial revolution, experts must have new skills and capabilities, learn different ways of communicating, master the materials one needs in one's field, and establish and diversify new working relationships with machines. What we need to note is that regardless of how AI develops, it is still being developed and that we humans should therefore be able to make good use of it.

---

3   Max Tegmark, *Life 3.0* (New York: Alfred A. Knopf, 2017).

History is related to a variety of academic fields, but the most founda-tion of historical research is historical records. Collecting and organizing historical records is the most fundamental and basic work of history. Cur-rently, many historical records are being digitized into images and texts. However, there are still some aspects that make it less easy for historians to use historical records. This paper intends to describe the difficulty of using historical texts recorded in Chinese characters as a source for pre-modern Korean history.

The first difficulty researchers face in using historical records is inter-pretation. Classical texts that are useful for studying Korean history are mostly written in Chinese characters. Chinese characters have different types, so even identical characters may differ depending on the type, re-quiring the ability to read Chinese characters by type. For example, *Sŭngjŏngwŏn ilgi* or *Hun'guktŭngnok* was written by a government agen-cy but is recorded in the cursive *ch'osŏ*. Personal letters and diaries that can be a major source of data in cultural history have also been recorded in *ch'osŏ*. It is not easy to read *ch'osŏ* because the handwriting varies depending on the writer. Also, since similar-looking characters may need to be understood as different characters depending on the context, the character can be understood only when one understands the context or the logic. Thus, interpreting the Chinese characters in the texts is itself a dif-ficult task.

The second difficulty is translating Chinese characters. Currently, we use Hangul as the main language, so the Chinese writing used in classical literature is archaic. So, in order to understand Chinese characters, one cannot read and use Chinese historical records unless one learns Chinese characters independently. In other words, the ability to translate Chinese characters is essential to using historical records. However, there are tem-poral and realistic difficulties in acquiring the ability to translate Chinese characters. The Institute for the Translation of Korean Classics (ITKC) states that it takes at least 10 years to train a translator who specializes in Chinese characters. Given that this is a professional translator, it takes more time and effort for someone specializing in another field, such as

history, to learn Chinese characters.

Third, there is the problem that a person can translate only a limited number of historical records in a given time frame. For example, the ITKC began to translate *Sŭngjŏngwŏn ilgi* in 1994. Twenty-three years later, in 2017, only about 20% has been translated. According to the ITKC, *Sŭngjŏngwŏn ilgi* consists of 1,841 books and is expected to take about 33 years to complete. The total number of Chinese classics at the ITKC to be compiled and translated is 18,483 books, including *Sŭngjŏngwŏn ilgi*, and this is expected to take 130 years.[4] There are many classical texts besides *Sŭngjŏngwŏn ilgi* that have yet to be translated *Sŭngjŏngwŏn ilg* . Translation also continues to require re-translation that reflects modern language, as the language currently used changes over time. So, while translation takes an enormous amount of time, the number of historical records one translator can translate is limited. Thus, it is expected to take even more time. Therefore, the reality is that the history researcher cannot simply wait for the historical record he or she wants to use to be translated.

Fourth, a variety of terms and background knowledge are required to translate historical records. For example, when royal and national events were held in the Chosŏn Dynasty (1392-1910), the overall details and drawings, including the discussions, procedures, and items to be prepared, were recorded and left as a document titled ŭigwe. The number of *ŭigwe* books that are currently available amounts to 3,869.[5] In the ŭigwe, unique terms appear in the lists of names of people or items, which can be understood only on the basis of diverse knowledge of geomancy, music, food, clothing, craft, and architecture. Also, because various forms of official documents were written in accordance with each format, there is a high probability of mistranslation if the translation is made solely based

---

4   Cho Seungrae(Sŭngnae), 2018. Press release for the parliamentary audit. October 12, 2018.

5   Seo Hanseok(Hansŏk), "Status of Euigwe Translation and Translation Methods," *National Culture* 42 (December 2013): 343-344

on the logic without considering the unique style and usage of official documents. Therefore, it has always been pointed out that the translation of ŭigwe   requires cooperative translation between different scholars or academic institutions.[6] Cooperation between many experts and academic institutions is needed not only for the translation of ŭigwe but also of other historical documents.

As such, the common problem of historical record translation is that individual scholars require reading ability of Chinese characters and relevant knowledge to translate classical texts, and this takes a lot of time and effort. This has been an obstacle to the expansion of historians' research horizons and to in-depth research. This has in turn limited the pace of progress in historical research.

The Current Status of Translating Historical Documents Using Artificial Intelligence

In the previous section, it was confirmed that in order to use classical texts in the course of historians' study, the difficulty of translation limited the use of the historical records by individual scholars. However, with the development of technology due to the fourth industrial revolution, the possibility of overcoming the limitations experienced by historians and expanding the quality and scope of historical research is expected to become more likely. Such attempts are already underway. What we want to pay particular attention to is the progress in translating Chinese texts using AI technology.

A major achievement is the progress of Korea's national project that can integrate AI technology into the translation of classical texts and translate historical records automatically, similar to Google Translate. The Ministry of Science and Information and Communications Technolo-

---

6 Lee Hyunjin(Hyŏnjin), "Difficulties in Translating the Euigwe - Focusing on Euigwe related to Hyungrye," *Classical Translation Studies* 5 (December 2014)

gy (ICT), which is in charge of the project, introduced new promising ICT such as the IoT, big data, and AI to the public sector, starting 2013, to foster innovation in public services.[7] ICT-based public service promotion projects have been carried out in order to solve social problems and contribute to the implementation of the intelligent information society.

As part of the project, the ITKC has been organizing efforts since 2017 to translate classical texts using AI. In 2017, it launched the "Building an Artificial Intelligence-Based Automatic Classical Text Translation System" with Chungnam National University, and in 2018 it launched the "Advanced Project for the Automatic Translation of Classical Texts Based on Artificial Intelligence." In 2019, the ITKC and the Korea Astronomy and Space Science Institute are working together to build the "Expanded Service Project for Cloud-based Classical Text Translation", which specializes in an area in automatic translation.

In 2017, the ITKC built an "AI-based Classical Text Automatic Translation System" that uses neural machine translation (NMT) technology in the translation of classical texts. The automatic translation system of classical texts was implemented by the ITKC using the technology of Systran, a leading company currently in charge of machine translation businesses worldwide. Based on a total of 350,000 corpora found in *Sŭngjŏngwŏn ilgi*, from Yŏngjo's year of enthronement to Yŏngjo 4 years and 3 months, provided by the ITKC for six months from June 2017, and based on the corpora that Systran possesses, the AI-based automatic translation model went through more than 30 training sessions. The neural machine translation technology applied to this translation system is one that finds unique patterns that are related to each other in the original and the translation and creates a translation model based on the data. This translation model learns information on the correspondence relationship between the original and the translation, and this allows it to explore and determine the

---

most appropriate word selection in the actual translation process before generating the final translation. At the same time, a tokenizer was developed for classical texts which separates classical Chinese characters into meaningful units and was incorporated into the system so that the AI translation engine could recognize the text more easily.[8] (Lee Ki-beom 2018).

In 2018, the ITKC started a project to upgrade the AI-based system for automatic translation of classical texts. Expanding the scope of the period for automatic translation to Injo's and Kojong's *Sŭngjŏngwŏn ilgi*, 430,000 classical text corpora were established in 2018. In addition, *Chosŏn wangjo sillok* was selected as a new translation target and the contents were extracted by subject from the time of the previous king to build the *Chosŏn wangjo sillok* corpus. Through this process, it promoted the enhancement of translation performance by expanding sentence patterns and eliminating non-registered words for automatic translation by securing various vocabulary words for each subject. By doing so, the sentence-length for machine learning was extended to 300 characters, from the initial, shorter sentence length of 150 characters or less, allowing smooth translation of even long sentences from the original. As such, in 2018, the ITKC started the task of generating tuned hyperparameters, an advanced translation model (see fig. 1), by strengthening the learning methods of the artificial neural network translation model.[9]

In order to understand the ITKC's automatic translation system, which is currently being promoted as a development and advancement project, it is necessary to understand the development of machine translation tech-

---

8 Lee Kibeom, "Systran Builds AI Classical Literature Automated Translation System," BLOTER. January 22, 2018. http://www.bloter.net/archives/300807 Accessed May 19, 2019.

9 ITKC (Institute for the Translation of Korean Classics), "2. Advancement of AI-based Automated Translation System for Classical Literature," *Public Service Promotion Project Briefing Data Sheet* (Seoul: ITKC, 2018): 39-52.
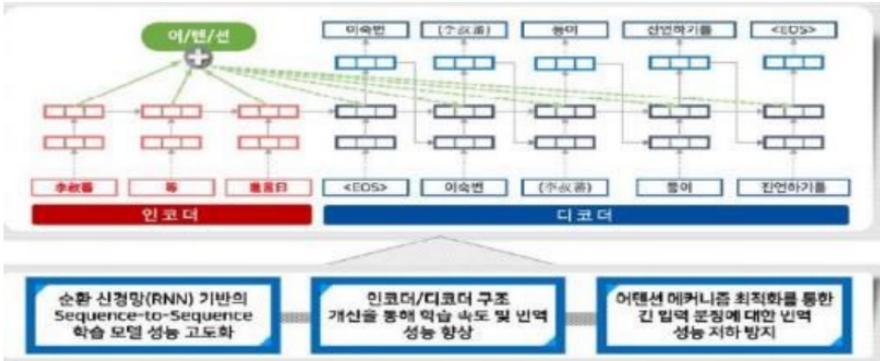
Fig. 1. Outline of the classical text application technology of the advanced learning model

nology. Machine translation has progressed through several stages before reaching the current level. The first is the Rule-based Machine Translation of the 1980s. This was a method in which a linguist or translator entered the language rules directly. Since humans had to make up the rules in machine translation, there was a limitation in creating a system that could translate a number of languages.

In the 1990s, Corpus-based Machine Translation was regarded as mainstream. In corpus-based machine translation, translation is done by gaining translation knowledge from a vast pool of single-words and target words. This is largely divided into Example-based Machine Translation (EBMT) and Statistical Machine Translation (SMT). EBMT is a method of translation by referring to examples most similar to input sentences in large target word corpora through the method of information search and sentence similarity measuring. The translated text and information are saved, and the translation results are given, using that information when there is a request for translation of the same sentence. In a limited corpus, however, there is a very low probability that there is a sentence that matches or is similar to an input statement.[10]

---

10 Shin Jisun(Chisŏn) and Kim Eunmi(ŭnmi), "Study on the Appearance of the Arti-

On the other hand, SMT dates back to the Cold War. At that time, Warren Weaver of the Rockefeller Foundation proposed to develop something similar to machine translation, and later, in 1988 IBM operationalized it for the first time.[11] SMT analyzes the linguistic correlation frequency between the original and translation, quantifies it, and uses it in translation. If sufficient parallel corpora for language pairs are provided, it statistically analyzes and models the linguistic characteristics that appear in the parallel corpus and automatically extracts the knowledge required in the translation even if you do not enter the rules for a given language's structure and characteristics. Since translation is to be done focusing on the statistical relation between the two languages instead of the linguistic background, more data increases accuracy based on statistical analysis. Since it is a translation method that does not require teaching the machine grammar and syntax, it can save the manpower, time, and cost required for developing a translation system for different language pairs when compared to the conventional rules-based translation system.[12] It is easier to translate with SMT than with rules-based translation since the system automatically learns the translation method. Also, it has the advantage that its translations are more complete.

The next technology that has recently drawn attention is Neural Machine Translation (NMT). With the announcement of Google's new automated translation system in the second half of 2016, AI has drawn attention among translation-related workers and researchers. It is a machine translation using deep learning algorithm technology that imitates human learning processes by using multi-layer neural networks. Unlike SMT,
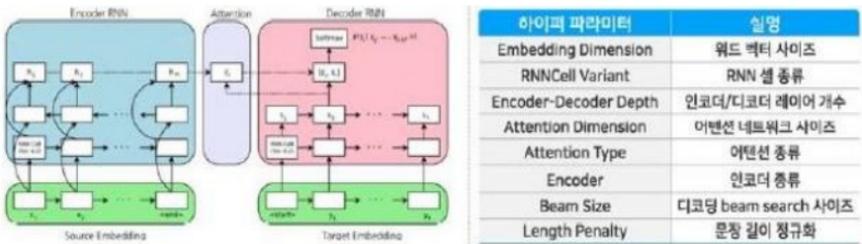
ficial Intelligence Translation System," *Translation Research* 18, no.5 (December 2017): 96-100.

11 Yoo Seongmin(Sŏngmin), "How far has machine translation evolved?" Science Times. October 11, 2018. https://www.sciencetimes.co.kr/?news=기계번역-어디까지-진화했나 Accessed May 19, 2019.

12 Shin Jisun and Kim Eunmi, "Study on the Appearance of the Artificial Intelligence Translation System," 97.

which translated words and phrases by splitting them up, NMT technology understands and translates sentences in their entirety. Since it can reflect word order, as well as the meaning and difference of context, comprehension of sentence context and accuracy are high. This technology is being applied to Google Translate and Papago.

In NMT, the computer reads the text by sentence unit and generates the best translation corresponding to the sentence using parameters acquired by deep learning. In general, NMT is based on the Recurrent Neural Network (RNN) algorithm. NMT is designed as a process of finding the optimum P(Y|X) by calculating the conditional probabilities of the input language (X) and output language (Y) using RNN. In NMT, an end-to-end method is used in which the input and output sentences are learned as a whole in translation.[13] Figure 2 is a schematic presentation of the NMT system being used to translate classical texts.

하이퍼 파라미터	설명
Embedding Dimension	워드 벡터 사이즈
RNNCell Variant	RNN 셀 종류
Encoder-Decoder Depth	인코더/디코더 레이어 개수
Attention Dimension	어텐션 네트워크 사이즈
Attention Type	어텐션 종류
Encoder	인코더 종류
Beam Size	디코딩 beam search 사이즈
Length Penalty	문장 길이 정규화

Source: ITKC 2018, 47.

Fig. 2. Outline of the hyper-parameter tuning of the classical texts automated translation system

NMT consists of three areas. On the left in figure 2 is the Encoder Network and on the right, the Decoder Network. There is an Attention module combined in the middle. Interpreting the input language (X) in an NMT system is generally called "encoder" and processing the output lan-

---

13 Kang Byungkyu(Pyŏnggyu) and Yi Jieun(Chiǔn), "The operating principle of neural network machine translation and the precision rate of translation--example of Chinese-Korean," *Chinese Literature* 73 (September 2018): 257-261, 286-287.

guage (Y) is called "decoder." The encoder is the part that reads the original text and decoder the part that generates the translation. In other words, the encoder identifies the information in the input sentence and uses RNN to read and store the information. The decoder is the process of finding the best translation for an input string based on information learned in parallel corpora.[14]

For example, if you enter "let's go to Starbucks" in the form of a natural language into the encoder, which is a code converter, the neural network encodes it into a digitized sentence using the embedded information of each word in the sequence. If you then put the encoded figure back into the decoder to perform the operation, it is finally decoded into a Korean sentence. A digitized sentence that has been encoded in the input phase is then entered back into the decoder at the output stage. For example, if "let's" is entered, the Korean word "가자," which is statistically the best translation word, is selected as an option and shown at the right. If "let's" and "go" are entered together, the probability for "가자" becomes even higher. In the same way, the probability that can be combined with "let's go" is calculated as "–로 가자" from the selected options, and when "to" is added, "–로 가자" is almost certainly selected. The calculation is completed when the proper noun "Starbucks," located at the end, is added. Predicting is a computational process in which probabilities are obtained based on internalized figures obtained through learning, and this is the working principle of AI translation.[15] The vector value of words or phrases that are thus mapped in space, forming relations, is called word embedding. This is the key technology of the new translation system.

Attention Mechanism is a method developed to efficiently handle long sentences in the course of NMT. The key point of "attention mechanism"

14 Kang Byungkyu(Pyŏnggyu) and Yi Jieun(Chiŭn), "The operating principle of neural network machine translation and the precision rate of translation--example of Chinese-Korean," 257-261, 286-287.

15 Lee Seungil(Sŭngil), "Discussing the Properties and Translators of AI Translation," *Translation and Interpretation Studies* 22, no.4 (November 2018): 194-195.

is to capture, weigh, and deliver information that should literally be viewed "with attention." The attention module reads both the information generated by the encoder and the information generated by the decoder and weights context information vectors that need to be viewed carefully to predict the language expression to be produced next. Since the RNN translation model encodes in fixed vectors, long sentences, as opposed to short ones, are difficult to encode with fixed vectors. As a result, long sentences are compressed into fixed vectors and processed, resulting in poor performance when translating them. Hence this problem is resolved using the attention mechanism.[16] The tuned automated translation system hyper-parameter, promoted by the ITKC in 2018, is similarly a neural machine translation system consisting of three elements: encoder, decoder, and attention.

Artificial neural networks produce high-accuracy translations based on these factors. The innovative feature of neural network-based methods is that first, the units of translation are expanded from words and phrases to sentences, enabling contextual recognition. Second, zero-shot translation technology allows AI to learn a language without learning data on its own. For example, if you teach it to translate from Korean to English and English to Japanese, AI can translate between Korean and Japanese at an appropriate level without going through an English translation.[17]

Table 1 compares the translation based on the parallel corpora and automated translation systems with that of a human being.

In the automated translation, the proper name 호랑(虎狼) was mistranslated as 호린(虎麟) and the verb 초복(初覆)에 입시하였다 was mistranslated as 초복했다, but it can be seen that the translation was smooth, reflecting the approximate flow of the context. The automated translation

16 Kang Byungkyu(Pyŏnggyu) and Yi Jieun(Chiŭn), "The operating principle of neural network machine translation and the precision rate of translation--example of Chinese-Korean," 260-261.

17 Shin Jisun(Chisŏn) and Kim Eunmi(ŭnmi), "Study on the Appearance of the Artificial Intelligence Translation System."

Table 1. Comparison of an automated translation and a human translation of *Sŭngjŏngwŏn ilgi*

【Yŏngjo 5th year (1729) November 26 (Byeongsin) 20th article】

又以義禁府言啓曰，　今月二十五日初覆入侍時，　捕廳罪人虎狼，　還發配所事，　榻前下敎矣。仍前罪目，　依例發遣府羅將，　押送于咸鏡道慶源府配所之意，　敢啓。傳曰，　知道。

Automated	Human (correct translation)
또 의금부의 말로 아뢰기를, 이달 25일 초복(初覆) 했을 때, 포도청의 죄인 호린(虎麟)을 도로 배소(配所)로 보내도록 탑전 하교(榻前下敎) 하셨습니다. 이전의 죄목에 따라 규례대로 본부의 나장(羅將) 을 보내 함경도 경원부(慶源府)의 배소(配所)에 압송하도록 감히 아룁니다. 전교하기를, 알았다.	또 의금부의 말로 아뢰기를, "이달 25일 초복(初覆)에 입시하였을 때 포도청 죄인 박호랑(朴虎狼)을 도로 배소로 보내도록 탑전 하교를 내었습니다. 이전 죄목 그대로 규례에 따라 본부의 나장(羅將)을 보내어 함경도 경원부(慶源府)의 배소로 압송하겠습니다. 감히 아룁니다." 하니, 알았다고 전교하였다.

Source: National Information Society Agency 2017.[18]

(demo ver 1.0) of *Sŭngjŏngwŏn ilgi* was evaluated by a translator at the ITKC and received an average score of 3 out of 5 points. In 2018, the score for the automatic translation results was further raised to 3.5 points based on the data for *Chosŏn wangjo sillok* and *Sŭngjŏngwŏn ilgi*.[19] This shows that AI translation can perform translations of Chinese historical records at a remarkable level.

If automated translation is used, it is expected that the translation of

---

18 Since the automated translation system currently under development at ITKC did not make their services available, I had to take advantage of translation examples provided by them. As will be mentioned later, the ITKC plans to provide services after 2020.

19 ITKC and Korea Astronomy and Space Science Institute, "20. Establishment of Cloud-Based Classical Texts Translation Service." *Public Service Promotion Project Briefing Data Sheet* (Seoul: ITKC, 2019), 259.

classical texts will become faster. The ITKC began translating *Sŭngjŏngwŏn ilgi* in 1994 and expected the translation to be completed in 2062. However, if deep learning-based AI translation is used, the translation is expected to be completed by 2035, 27 years sooner than the initial 45 years.[20]

Also, in 2019, the ITKC has started a translation project of classical texts in special fields. At present, it is building a platform that will open up and expand its know-how in AI-based classical text translation in order to digitize ancient astronomical records for use in the field of astronomy. Since the translation method and difficulty depend on the type of text, translation accuracy can be increased if a specialized machine translation system is developed for each text type or field and learns the formalized language pattern used in that field. Since specializing by text type increases machine translation accuracy, it is necessary to develop general translation functions and domain specific translation functions selectively.[21] The human translation of classical texts in specialized fields was of low quality because experts in that field lacked reading skills in Chinese characters and translators specializing in classical texts had insufficient knowledge of the various fields. The ITKC plans to gradually expand the number of participating institutions to provide a comprehensive range of automated translation models, and to develop and implement a specialized classical text translation service tailored to the participating agencies. Thus, users of classical texts can improve the accuracy of the translation by selecting and using a model that matches the characteristics of the classical text that they want to translate (ITKC and Korea Astronomy and Space Science Institute 2019, 254-263).[22]

---

20 Ministry of Science and ICT, *Korea Information Society Agency, Public Service Promotion Project Briefing Data Sheet* (Sejong: Ministry of Science and ICT, 2018), 10.

21 Kang Byungkyu(Pyŏnggyu) and Yi Jieun(Chiŭn), "The operating principle of neural network machine translation and the precision rate of translation--example of Chinese-Korean."

22 ITKC and Korea Astronomy and Space Science Institute, "20. Establishment of
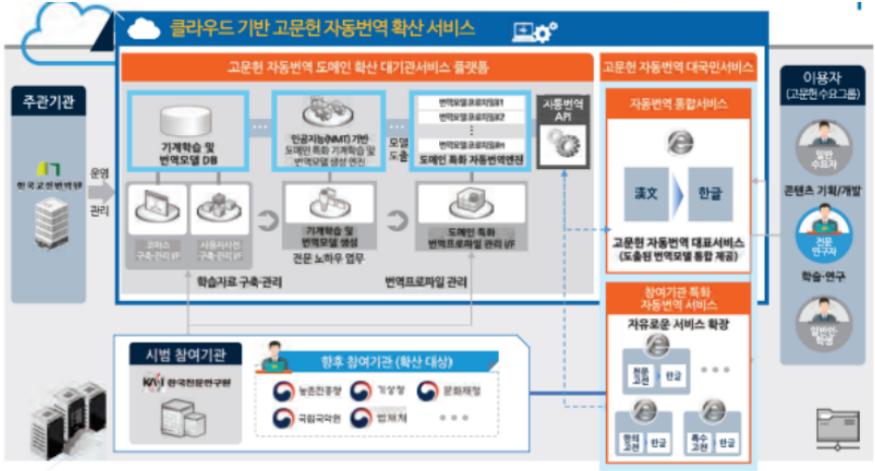
Source: ITKC and Korea Astronomy and Space Science Institute 2019, 258.

Fig. 3. Conceptual map of the expansion and platform of cloud-based classical text automated translation

According to the ITKC, it will make an automated translation service for classical texts available to the public after 2020 (ITKC and Korea Astronomy and Space Science Institute 2019, 51). Although the translation system is at the level of a first draft, AI translation will reach a more sophisticated level as more time goes into building it and more historical records are digitized. Also, it will be able to accumulate information from various fields far beyond the amount of information any individual can hold. Therefore, an individual scholar will be able to translate historical records without the time, physical effort, and processes required for interdisciplinary exchange.

Thus, if we use automated translation using AI, it is expected that historians will be able to carry out more productive historical research as the time required for the translation of historical records is shortened so that more time can be spent on creative research.

Cloud-Based Classical Texts Translation Service."

The Limitations of Artificial Intelligence Translation
and the Role of Historians

In the age of the fourth industrial revolution, the translation of classical texts through AI will soon be a breakthrough mechanism for people to overcome the limitations of using historical records written in the Chinese language. Historians will be able to actively use the convenience of automated translation provided by AI in their research. If the translation of vast quantities of historical records is carried forward with AI, historians will have access to more classical texts with less time and effort. In other words, access to classical texts will be further improved.

However, the benefits of increased access to classical texts are not limited to historians alone. Even the public will be able to read and understand classical texts without the ability to translate Chinese characters. Already, many historical records have been digitized and are being opened to everyone, and this trend will further accelerate and expand the sharing of knowledge in the future. Previously, historians were able to gain authority as scholars simply by approaching historical records that were not accessible to the general public. But now, as historical records are gradually becoming digitized and translated using AI and other technologies, the authority and profile that historians could gain from monopolizing historical records will disappear. This means that understanding and thinking about history has become an open subject for the various masses to share and discuss, instead of something exclusively carried out by scholars in academia. This will further narrow the gap between historians and the members of the public who enjoy history and personally study it. In particular, the academic realm of history, which pursues history as a fact of the past, will not be the sole preserve of historians anymore. Therefore, it can be asked whether historians will have a role to play in the future and whether the academic realm of historians is unique, original, or essential. What is the role and expertise of historians as scholars who can have a distinct identity from the public? On what basis will they demonstrate the value and role of being a historian?

On the other hand, historians must also compete with cutting-edge machines, including AI. In the age of the fourth industrial revolution, that human distinction by which man can survive among machines is called creativity. Creativity is an ability that only humans can display because AI and machines do not come up with creative ideas but only perform according to human-inserted computations. Creative ideas and free imagination are unique abilities that only humans are capable of and AI cannot keep up with. These are expected to be used more significantly in future work. Thus, historians will also be required to exercise creativity. Historians should be able to do creative historical research that is different from the translations that machines or AI perform. The historian him- or herself should be able to actively look at history and gain insight as the agent of interpretation, as a being different from machines. This means that historians should no longer play the simple role of organizing and introducing historical materials or analyzing past facts. More and more historians will now be asked to conduct a more insightful and creative study of history.

What is the source of creative thinking for historians? On what basis can the historian interpret history in a way that distinguishes him or her from the general public or a machine? There may be a number of ways for historians to think creatively, but in relation to the theme of this article, the translation of classical texts, historians who use classical texts need an improved understanding and learning of Chinese characters.

The claim that Chinese characters need to be learned seems to contradict the prospect that the development of AI will enable historians to overcome the limitations they face in the translation of classical texts. To understand this, we need to take a closer look at the translation of classical texts. In the future, AI will complete the translation of classical texts to a considerable degree. However, human use of language is not only realized through the mechanical application of language grammar, but is constantly created and adapted based on creativity, such as is done when drawing pictures and making music. If creativity is a human ability that AI cannot acquire, it will not be able to understand and translate all aspects of linguistic activity, including the use of language and the play on

words.

Above all, there is a limit to the NMT that is being used in the translation of classical texts in the present.[23] First, NMTs take a lot of time to learn. Second, AI translation is possible only if the text is typed into a computer. However, not all of the classical texts translated are digitized.

Third, data from Chinese historical records do not meet the conditions for good translation using NMT, unlike the languages we currently use. It is not easy to produce a good translation with NMT with a small amount of corpus data. At least millions of sentences are needed for universal NMT to work smoothly. NMT performs well in an environment where big data-level translation data is secured. As mentioned earlier, however, the classical texts that have been translated are only a small part of the entire classics. If we consider the data being used by the classical texts automated translation system developed by the ITKC, *Sŭngjŏngwŏn ilgi* was once limited to the time of Yŏngjo but now includes that of Injo and Kojong, and also uses *Chosŏn wangjo sillok*. The total number of letters in *Chosŏn wangjo sillok* is 49,646,667, while the remaining letters in *Sŭngjŏngwŏn ilgi* are about 242,500,000. In terms of the number of letters alone, one Chinese character is usually two bytes. Thus, *Chosŏn wangjo sillok* is 94.7 MB and *Sŭngjŏngwŏn ilgi* is 462.5 MB, all falling short of 1 GB. However, global data generated annually through social network and media is measured in zettabytes (ZB). 1 ZB is 1.1 trillion GB. Compared to this, less than 1 GB of data from *Chosŏn wangjo sillok* and *Sŭngjŏngwŏn ilgi* is a small amount for AI to learn from. Furthermore, *Sŭngjŏngwŏn ilgi* and *Chosŏn wangjo sillok* are government-created materials and are historical records written in highly refined Chinese sen-

---

23 Seo Bohyun and Kim Soonyoung, "A Review of Error Types in Machine Translation Results," *Translation Research* 19, no.1 (March 2018); Kwon Do-kyung et al, "The Status of Chinese-Korean Translations by AI Translators in 2018 – Focusing on Google and Papago Translation," Korea Chinese Literature Society Academic Conference Collection 2018; Ma Seunghye, "Comparison, Review and Discussion on Human Ability and Literary Translation Difficult to Mechanize," *Translation and Interpretation Studies* 21, no.3 (August 2017).

tences. However, there is a limit to using the *Chosŏn wangjo sillok* and *Sŭngjŏngwŏn ilgi* as data for translations of classical texts, which vary depending on various writing styles and formats, including Chinese poems about ancient events or using characters that are implicit in meaning, or ancient texts based on Chinese transcription.

Fourth, since there are significant mistranslations in existing translations and each translator has room to interpret the same phrase differently, the standard of the texts that are used as data is not very high. Since there is no space between Chinese characters, AI has to sort out words on its own. But since Chinese characters are ideograms, there is no rule for spacing and the meaning of the sentence varies greatly depending on the spacing. Thus, different translators mark the text differently. In some cases, the word order of Chinese characters is applied differently; one letter may have many meanings; and different readings may be required in the same context. Thus, the translation is performed differently depending on the translator's literary and technical skills. It is questionable how much AI can overcome such differences resulting from different translators. So, although AI may be capable of rapid translation, it will be able to translate Chinese sentences that are both implicit and anomalous only when supported by a great amount of time and high technological advancement.

If language is the result of creative and inventive production activities performed by humans, it should be recognized that there is a limitation to machine translation. That is, although translation using machines may be done quickly, it is the responsibility of humans to check for errors that the machine did not recognize or for misinterpretations. The ability to discover such errors will become a requirement for experts. In that sense, historians should not rely solely on AI translation techniques but should be able to actively translate based on historical knowledge, recognize differences in translation, and discover errors in translation. This would require of a historian also to be an expert linguistic researcher.

On the other hand, a historian should accumulate a variety of knowledge and information in him- or herself to interpret history since he or she has a point of view and a worldview from which to capture history

and from which to read it. Such knowledge and information are not obtained solely by the acquisition of historical facts. These include areas of human emotions, thoughts, and sensibilities that machines cannot feel. The agent that moves history is the human being, the subject that explores and understands history is the human being, and human emotions and affections captured in the course of history are something that only human beings can understand. Language is a useful tool and material for understanding the human mind and culture inherent in such a trend of history.

Chinese characters originated from the ancient inscriptions of Shang, traversed China and arrived in Joseon, being formed, used, and transformed in a long history. Thus, the usage and format of Chinese characters contain the mindset and intelligence of the people of the time who used them. Therefore, Chinese characters are the product of the total civilization of time and creative work. So Chinese characters are not only language symbols that interpret historical materials but another channel that connects us with the past. When translating Chinese characters and exploring them, historians will be able to experience the culture, thought, and the human emotions and affections that are contained therein.

In order to get closer to the facts of the past, as well as to experience people's way of thinking, their culture and experience, historians must learn Chinese, the language of the past and a total collection of cultures and times. By embodying it deep inside oneself, one can better understand and look at the human and their lives in the past. Unlike the general public's approach and interpretation of historical records in which historical facts are discovered and reviewed, the historian will be able to understand the times and look at the times with insight based on the profound viewpoints that historical records provide.

In the era of the fourth industrial revolution, historians need to speed up the pace of research by making full use of the classical text translations provided by machines to provide access to a wider range of historical texts. They should continue to understand and learn Chinese characters as scholars without becoming complacent in the convenience of translation provided by machines. By learning Chinese characters which are not used

in modern times, one should slowly immerse oneself in the way people thought in the past. Instead of being limited to the translation possibilities and limitations provided by machine translation, it is hoped that historians would attempt translations of historical texts independently and move on to more creative historical explorations.

Conclusion

This study began with the awareness that historians need to recognize the importance of change and that historical studies must cope with these future changes in the era of the fourth industrial revolution, in which human life will undergo radical and fundamental changes with new technology represented by AI, robotics and the IoT. Therefore, it looked at how historians can use AI in historical research and their role as historians in the translation of classical texts which is the most basic work of historical research.

Historians have had a hard time translating classical texts. There were difficulties in reading Chinese characters and acquiring the ability to translate them, the quantity of classical texts was vast, and knowledge in various fields had to be used. This has limited the pace of development in historical research. However, in the era of the fourth industrial revolution, an automated translation system for classical texts was built with AI. If more data and time are invested, a higher level of translation using AI is expected. Based on this, it is expected that historians will be able to carry out more productive historical research as the time and effort spent on translating classical texts is saved and more time can be invested in creative research.

Nevertheless, we should not become complacent with the convenience of translation provided by machines in the era of the fourth industrial revolution but learn Chinese characters and further acquire the ability to translate them. With automatic translation, the public will also find easier access to historical records, further narrowing the gap between historians

and the public in terms of the simple role of examining past facts. Historians will also be required to carry out historical studies in which creativity, the unique ability that humans but not machines have, is more actively exercised. Therefore, the historian should be able to make a more creative and insightful quest for history by learning Chinese characters, a total collection of long-standing human civilization, and by experiencing the minds and cultures of past people. Also, AI will not be able to translate classical texts completely, fully understanding humans' creative use of language. By learning Chinese characters, they should have the expertise to independently translate these and check for errors in the translations.

As such, historians will have greater access to and use of classical texts by using AI automated translation, further expanding the scope of their research. Nevertheless, it will be necessary to continue learning Chinese characters rather than relying solely on the translations provided by technology, so that they can exercise their expertise and creativity in research, a quality that will be even more demanded of historians in the era of the fourth industrial revolution.

Whether AI will provide greater opportunities for the future of history or guide it to the path of regression is up to us to think about and make choices. Historians should build greater possibilities and opportunities for the future of historical studies through serious thought and reflection.

References

1. Cho, Seungrae(Sŭngnae). Press release for the parliamentary audit. October 12, 2018.
2. ITKC (Institute for the Translation of Korean Classics). "2. Advancement of AI-based Automated Translation System for Classical Literature." *In 2018 Public Service Promotion Project Briefing Data Sheet*. Seoul: ITKC, 2018.
3. ITKC and Korea Astronomy and Space Science Institute. "20. Establishment of Cloud-Based Classical Texts Translation Service." In

2019 Public Service Promotion Project Briefing Data Sheet. Seoul: ITKC, 2019.

4. Kang, Byungkyu(Pyŏnggyu) and Yi Jieun(Chiŭn). "The operating principle of neural network machine translation and the precision rate of translation--example of Chinese-Korean." *Chinese Literature* 73 (September 2018): 257-261, 286-287.

5. Kim, Soyoung(Soyŏng) et al. *The Ghost of the Fourth Industrial Revolution*. Seoul: Humanist, 2017.

6. Kwon, Dokyung et al. "The Status of Chinese-Korean Translations by AI Translators in 2018 – Focusing on Google and Papago Translation." Korea Chinese Literature Society Academic Conference Collection 2018.

7. Lee Hyunjin(Hyŏnjin). "Difficulties in Translating the Euigwe - Focusing on Euigwe related to Hyungrye." *Classical Translation Studies* 5 (December 2014).

8. Lee, Kibeom. "Systran Builds AI Classical Literature Automated Translation System." BLOTER. January 22, 2018. http://www.bloter.net/archives/300807 Accessed May 19, 2019.

9. Lee, Seungil(Sŭngil). "Discussing the Properties and Translators of AI Translation." *Translation and Interpretation Studies* 22, no.4 (November 2018): 194-195.

10. Ma, Seunghye. "Comparison, Review and Discussion on Human Ability and Literary Translation Difficult to Mechanize," *Translation and Interpretation Studies* 21, no.3 (August 2017).

11. Ministry of Science and ICT. *Korea Information Society Agency, Public Service Promotion Project Briefing Data Sheet*. Sejong: Ministry of Science and ICT, 2018.

12. Nam, Chuljin(Ch'ŏljin). "The Problems of Google Translation (GNMT) Viewed Through Chinese-Korean Translation." *China and Chinese Studies* 34 (May 2018).

13. National Information Society Agency. "Press material on classical texts translation system." December 20, 2017.

14. Schwab, Klaus. *The Fourth Industrial Revolution*. London: Penguin

Press, 2017.

15. Seo, Bohyun and Kim Soonyoung. "A Review of Error Types in Machine Translation Results." *Translation Research* 19. no. 1 (March 2018).

16. Seo, Hanseok(Hansŏk). "Status of Euigwe Translation and Translation Methods." *National Culture* 42 (December 2013): 343-344

17. Shin, Jisun and Kim Eunmi. "Study on the Appearance of the Artificial Intelligence Translation System." *Translation Research* 18, no.5 (December 2017): 96-100.

18. Tegmark, Max. *Life 3.0*. New York: Alfred A. Knopf, 2017.

19. Yoo, Seongmin(Sŏngmin). 2018. "How far has machine translation evolved?" Science Times. October 11, 2018. https://www.sciencetimes.c o.kr/?news=기계번역-어디까지-진화했나 Accessed May 19, 2019.

⟨Abstract⟩

The Translation of Historical Documents and the Study of Korean History Using Artificial Intelligence

Geunhye Hong

The so-called fourth industrial revolution promises to change human life radically with new technology, including Artificial Intelligence. The translation of ancient Chinese texts as a source of knowledge about pre-modern Korean history is a challenge for historians because this requires a mastery of Chinese and its ancient characters, which is hard to acquire. This article describes how Artificial Intelligence has been used since the 1980s to create translations and how it can be used to overcome the translation challenges that Korean historians face. It explains the work that the Institute for the Translation of Korean Classics has been doing to use Artificial Intelligence for the translation of historical Chinese texts, as well the current status of the work. The article concludes by discussing the limitations of Artificial Intelligence when translating ancient Chinese into Korean and attempts to define a new role for historians alongside such translation programs.

Keywords: Artificial Intelligence, Translation, Ancient Chinese, Korean History, Historians

〈국문초록〉

인공지능을 통한 고전문헌 사료 번역과 역사학의 미래

홍근혜 (고려대학교 박사수료)

본고는 인공지능, 로봇공학, 사물인터넷 등으로 대표되는 새로운 과학기술로 인류의 삶이 급진적이고 근본적인 변화를 겪게 될 4차 산업혁명의 시대에 역사학자가 변화의 중요성을 인지하고, 역사학이 이러한 미래의 변화에 대처해야한다는 문제의식에서 출발하였다. 그리하여 역사 연구의 가장 기본 과정인 고전문헌 사료 번역에서 역사학자들이 인공지능 기술을 연구에 활용하는 방안과 역사학자로서의 역할을 고찰해 보았다.

인공지능 기술을 활용한 고전문헌 자동번역 시스템이 구축되어 인공지능이 고전문헌을 번역할 수 있게 되었다. 앞으로 인공지능을 활용한 더 수준 높은 고전문헌 번역이 이뤄질 것으로 예상된다. 이를 활용함으로써 역사학자들은 사료 번역에 소모되는 시간과 노력을 단축하고 더욱 생산적인 역사 연구를 수행할 수 있을 것으로 기대된다.

그럼에도 불구하고 4차 산업혁명 시대에 역사학자는 한문을 학습함으로써 전문성과 창의성을 발휘하는 연구를 수행할 수 있어야 한다. 한문 자체에 내재해 있는 과거 사람들의 심성과 문화를 내면에 체감하여 더욱 창의적이고 통찰력 있는 역사 탐구를 할 수 있을 것이다. 또한 인공지능이 인간의 창조적인 언어를 이해하여 고전문헌을 완벽하게 번역할 수는 없으므로 한문을 독자적으로 번역하여 자동번역의 오류를 확인할 수 있는 전문성을 갖추어야 한다.

인공지능이 역사학의 미래에 더 큰 기회를 제공할 것인지 아니면 역사학을 퇴보의 길로 안내할 것인지는 우리가 고민하고 선택해야 한다. 역사학자들은 진지한 고민과 성찰을 통해 미래의 역사학에 더 큰 가능성과 기회를 구축해 나가야 할 것이다.

주제어: 인공지능, 4차 산업혁명, 번역, 고전문헌, 역사학