Toward a More Accurate and Easy-to-use System for Rapidly Evolving Machine Translation

Masaaki Nagata
Senior Distinguished Researcher, NTT Communication Science Laboratories

Abstract

Chat Generative Pre-trained Transformer (ChatGPT), a generative artificial intelligence (AI) chatbot developed by OpenAI, is rapidly gaining attention worldwide. ChatGPT generates sentences as a response to input sentences or words and outputs the sentences in the language of the input sentences unless otherwise specified. However, it can be used for machine translation; that is, it outputs a translated sentence when the output language is indicated. Masaaki Nagata, a senior distinguished researcher at NTT Communication Science Laboratories, has been researching natural-language processing and its application, machine translation, for more than 20 years. We asked him about the trends and characteristics of translation using large language models (LLMs) and machine translation using a Japanese-English bilingual patent corpus, which is about to be commercialized. We also elicited his thoughts on the research process and ideas as being a result of encounters.

Keywords: machine translation, bilingual corpus, large language model

Machine translation has evolved from statistical machine translation to neural machine translation and translation using LLMs

—Could you tell us about the research you are currently involved in?

I'm researching—in the field of machine translation—a Japanese-English bilingual corpus and word alignment in that corpus.

In the previous interview in this journal (June 2021 issue), I talked about the period of transition from statistical machine translation to neural machine translation. With neural machine translation, which has been rapidly gaining in popularity over the past few years, it is important to collect bilingual data. Therefore, we created bilingual patent data based on our experience of creating a Japanese-English bilingual corpus called JParaCrawl containing over 20 million sentence pairs through web crawling. Patents are public documents, and all documents of patent applications in Japan, USA, and other countries are publicly available. Since we had the expertise to create large-scale bilingual data, we were able to use it to create a Japanese-English bilingual patent corpus with over 300 million sentence pairs (**Fig. 1**).

One possible application of this corpus is creating and checking related documents when applying for

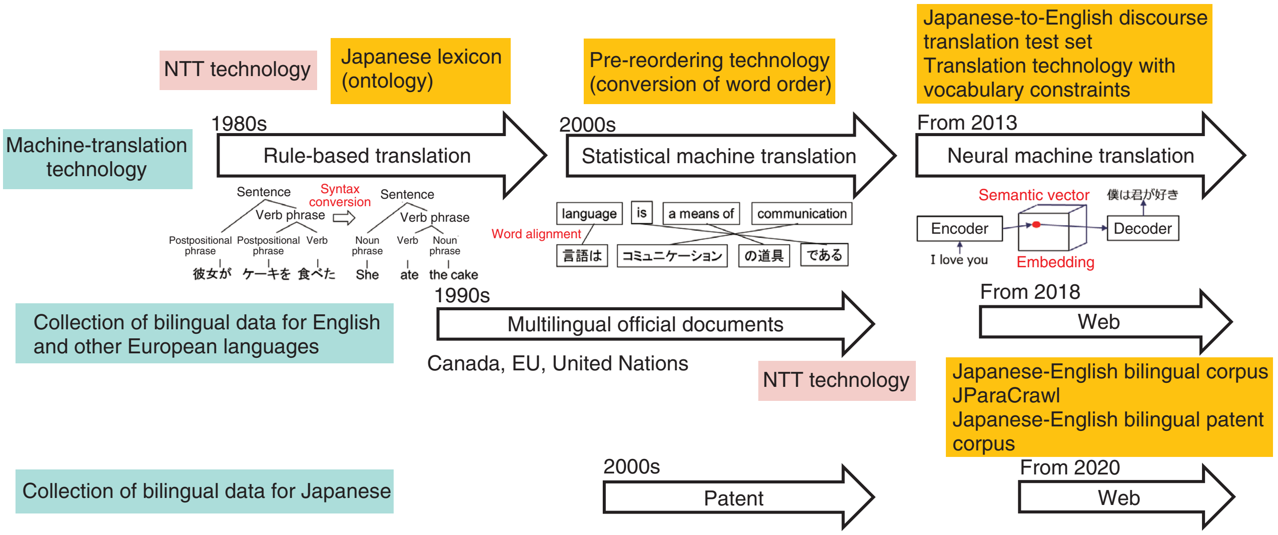
NTT technology
Japanese lexicon (ontology)
Pre-reordering technology (conversion of word order)
Japanese-to-English discourse translation test set
Translation technology with vocabulary constraints
Machine-translation technology
1980s
Rule-based translation
2000s
Statistical machine translation
From 2013
Neural machine translation
Sentence
Syntax conversion
Sentence
Verb phrase
Verb phrase
Postpositional phrase
Postpositional phrase
Verb
Noun phrase
Verb
Noun phrase
彼女が
ケーキを
食べた
She
ate
the cake
language
is
a means of
communication
Word alignment
言語は
コミュニケーション
の道具
である
Semantic vector
僕は君が好き
Encoder
Decoder
I love you
Embedding
Collection of bilingual data for English and other European languages
1990s
Multilingual official documents
Canada, EU, United Nations
From 2018
Web
NTT technology
Japanese-English bilingual corpus
JParaCrawl
Japanese-English bilingual patent corpus
Collection of bilingual data for Japanese
2000s
Patent
From 2020
Web

Fig. 1. History of machine translation and research and development at NTT.

patents overseas. For translation of patent-related documents such as patent applications prepared when applying for a patent overseas, it is necessary to maintain the relationship between a modification and the modified subject and to select appropriate technical terminology to secure the necessary rights. Logical precision and rigor are required, even if it means sacrificing some level of fluency, which is important in general translation. To evaluate if our corpus satisfies these requirements, we ranked the translation results of three types of artificial intelligence (AI)-based translation engines: (i) a patent-specific translation engine built using our Japanese-English bilingual patent corpus, (ii) a general-purpose translation engine, and (iii) a patent-specific translation engine by using Japanese and English patent applications of Mitsubishi Heavy Industries (including claims) as evaluation data. The quality of translation was ranked by employees working on patent applications in the intellectual-property departments of both Mitsubishi Heavy Industries and NTT. The average rankings of the three engines were respectively 1.5, 2.0, and 1.9, which indicates that the highest ranking was achieved by the patent-specific translation engine built using our bilingual patent corpus. An automatic evaluation based on similarity to the correct translation prepared in advance by a professional translator showed a significant improvement in translation by the patent-specific translation engine built using our bilingual patent corpus (scoring 57.5 out of 100 points) compared with 38.6 points for the general-purpose translation engine and 44.0 points for the other patent-specific translation engine. Mirai Translate, an NTT Group company, plans to provide a patent-specific translation engine built using our bilingual patent corpus as a service.

Although the accuracy of machine translation has increased considerably, it tends to decrease as sentences become longer, as in the case of patent claims. Our recent research theme is to clarify whether it is possible to improve the accuracy of translation of long sentences by using the world's most accurate word-alignment technology that we have created or to create a system that automatically identifies mistranslations or potential mistranslations. To improve translation accuracy in fields other than patents, we are considering the use of large language models (LLMs), which have been gaining popularity over the past year.

—How do you use an LLM for translation?

To put it simply, machine translation using an LLM is like asking a multilingual speaker who has learned a lot of things in English to paraphrase sentences in one language into another language. To understand the difference between translation using an LLM and conventional neural machine translation, it is necessary to look at the history from neural machine translation, to transformers, to the appearance of LLMs. In neural machine translation, the meaning of a sentence is represented by a real-number vector with about

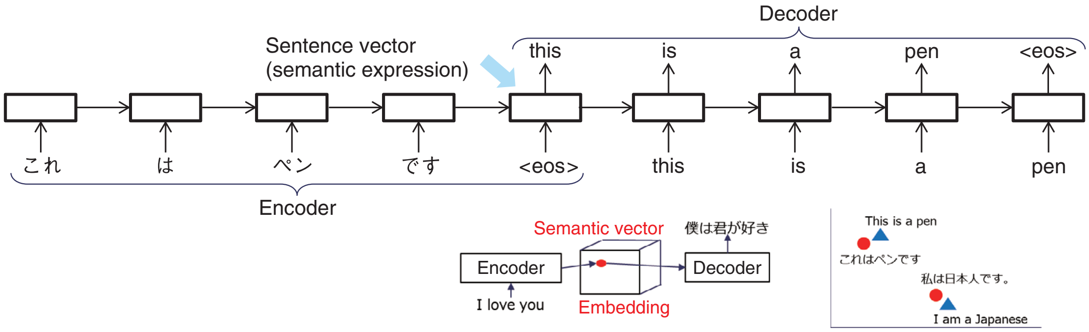

Fig. 2. Neural machine translation (RNN encoder-decoder model).

1000 dimensions then translated by coupling an encoder, which converts an input sentence in one language into a numeric vector, with a decoder, which converts this numeric vector into an output sentence in another language. In regard to the early version of neural machine translation, an encoder and decoder are configured as a recurrent neural network (RNN) into which words are input one by one, the internal state of the RNN is updated, and words are output one by one according to the internal state (**Fig. 2**). The internal state of the RNN when the special symbol <eos> (representing the end of the input sentence) is input contains information about all the words in the input sentence, so it is treated as a semantic representation of the input sentence (semantic vector). Training this RNN encoder-decoder model with a large amount of bilingual data will, for example, place the semantic vectors of the sentence pair “this is a pen” in English and its translation in Japanese (“kore wa pen desu”) in close proximity. In other words, the similarity of sentences corresponds to the distance between semantic vectors in vector space. Neural machine translation translates by sharing one vector space between two languages in this manner.

Encoders and decoders using an RNN are afflicted with two problems. First, they represent input sentences as fixed-length vectors, so long sentences are difficult to represent; second, they input words one by one and output words one by one, so parallelization is difficult. To solve these problems, a neural network called a transformer was devised (**Fig. 3**), which uses an attention mechanism to selectively collect information from multiple targets. A transformer is composed of an encoder, which selects information from all the words in the input sentence and reconstructs each word vector of the input sentence, and a decoder, which selects information from each word vector of the reconstructed input sentence as well as all the word vectors in the previous output and determines the next word to output. A transformer is more accurate than an RNN encoder-decoder model, and the attention mechanism allows for parallelization, which makes it possible to train translation models on large amounts of bilingual data and to scale the translation models. Our Japanese-English bilingual patent corpus also uses a transformer to create translation models.

An LLM called Generative Pre-trained Transformer (GPT)-3, on which OpenAI’s ChatGPT is based, has been attracting attention. As its name suggests, GPT-3 is a transformer model but consists of only a decoder, and the model is scaled and trained with a large amount of text data. The decoder-only transformer model—called a language-generation model or simply language model—is a neural network that uses an attention mechanism to select information from all the previously output word vectors and determines the next word to output. For GPT-3, a model with 96 layers and 175 billion parameters is trained with text composed of 0.3 tera tokens. It is called an LLM because it is approximately 1000 times larger than previous translation models and language models. ChatGPT, an interactive chatbot based on GPT-3, can be used comfortably in Japanese, and for some reason, it can also translate Japanese to other languages. However, most of the training data used for GPT-3 is in English, and Japanese accounts for only about 0.2% of the total data. Although it is not yet clear why ChatGPT can be used comfortably in languages other than English, it is probably due to the fact that all the languages included in the GPT-3 training data share one semantic vector space at a fine-grained level such as words (**Fig. 4**).

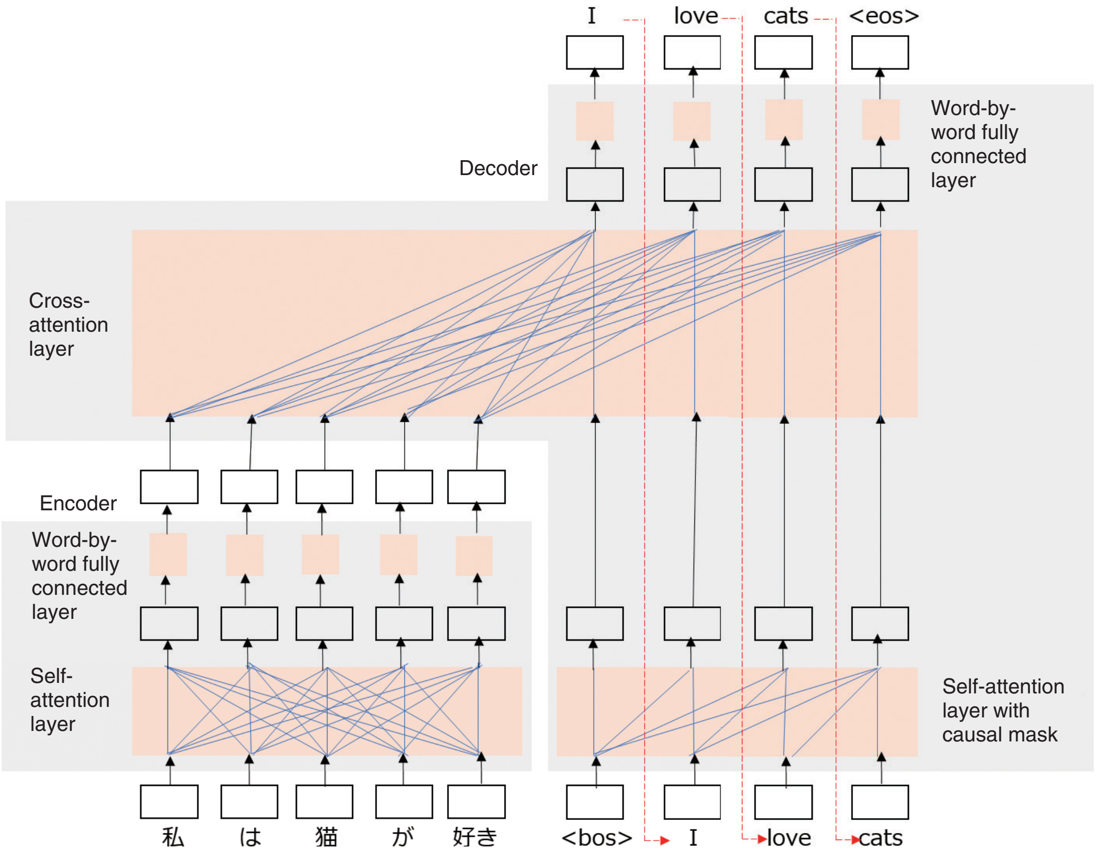
I
love
cats
<eos>
Decoder
Word-by-word fully connected layer
Cross-attention layer
Encoder
Word-by-word fully connected layer
Self-attention layer
Self-attention layer with causal mask
私
は
猫
が
好き
<bos>
I
love
cats

Fig. 3. Neural machine translation (transformer).

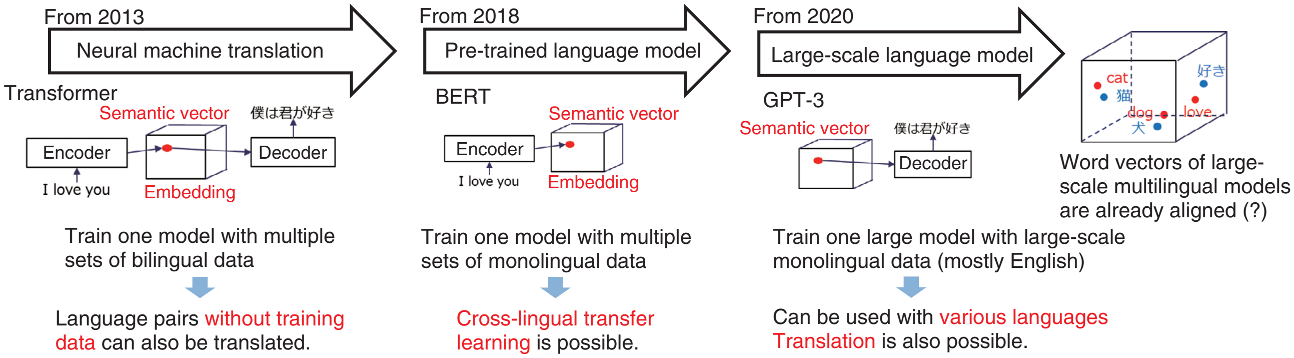
From 2013
Neural machine translation
Transformer
Semantic vector
僕は君が好き
Encoder
Decoder
I love you
Embedding
Train one model with multiple sets of bilingual data
Language pairs without training data can also be translated.
From 2018
Pre-trained language model
BERT
Semantic vector
Encoder
I love you
Embedding
Train one model with multiple sets of monolingual data
Cross-lingual transfer learning is possible.
From 2020
Large-scale language model
GPT-3
Semantic vector
僕は君が好き
Decoder
Train one large model with large-scale monolingual data (mostly English)
Can be used with various languages
Translation is also possible.
cat
猫
dog
犬
好き
love
Word vectors of large-scale multilingual models are already aligned (?)

Fig. 4. Multilinguality of language models.

LLM translation can be likened to, for example, a bilingual person reflexively speaking English after hearing something in Japanese.

—ChatGPT is evolving rapidly, but what does the future hold for machine translation?

At the present stage, translation accuracy of machine translation using LLMs is equivalent to that of conventional neural machine translation; however, compared with neural machine translation, LLM-based machine translation uses 1000 times more parameters, so the processing is naturally slower. Due to its large scale, it is also difficult to prepare and operate a system for LLM-based machine translation; thus, I predict that conventional neural machine translation will not disappear. However, it is also true that translations using ChatGPT are popular with users. I think this popularity is due to the fact that ChatGPT can be controlled using natural language and can do things such as consider longer contexts, handle specified style and terminology, and convert input/output formats.

Under these circumstances, research topics concerning machine translation include evaluation of the quality of machine translation using LLMs, detection and correction of errors, confirmation of factuality, and detection of biases such as gender bias. The current trend of creating huge, supreme intelligence is expected to reach its limit at some point. I therefore believe that our future challenge will be to (i) create instruction-tuning data that reflects the usage of machine-translation systems by professional translators and (ii) develop a smaller LLM "translation assistant" with around 10 billion parameters that enables users to instruct machine translation and its peripheral functions in natural language.

The research process and ideas are a result of encounters. Increase the chances of encounters through trial and error

—What do you keep in mind as a researcher?

In my experience, I feel that much of the research I have conducted while aiming at achieving a result presumed in advance has not been successful. Some things I tried led to new results. For example, for the aforementioned LLM-based translation, fine-tuning the LLM with ordinary bilingual data did not work; however, through trial and error, I found a pattern that worked. Such a good process and good ideas are a result of encounters.

Therefore, to have many chances for such encounters, I value meeting with people, including those involved in joint research, and trial and error. In joint research, we try to avoid overlapping work as much as possible and repeat trial and error without preconceptions; however, each researcher may do what they want, and sometimes that leads to good results. I also think that many encounters occur during periods of technological innovation. During periods of innovation, little established practice exists, so it is inevitable that people do a lot of trial and error. With the advent of LLMs, we are currently in the midst of a period of innovation, so I want to focus on trial and error from the perspective of translation.

—What is your message to younger researchers?

My original research theme was natural-language processing, but for the past 20 years, I have been working on themes related to machine translation, which is one type of natural-language processing. I enjoy learning foreign languages, so this field is one that I am interested in. Basic research involves a long-term endeavor, so unless you can maintain your interest in it, you will not be able to work on it for a long time. If you research in the same field for a long time, the external factors surrounding your research will change from time to time, so you will have to study to keep up with those changes; however, by keeping your purpose and goal in mind, you can move forward without losing sight of yourself.

In the previous interview, I said something like "do something different from others and be as controversial as possible in a good way." I'm not saying one should be different from others. As researchers, we must always be aware of—even within the same field—tackling a new research problem, gaining new knowledge, and creating new technologies, which mean doing something different from others. To tackle a new research problem, you have to understand the current situation, study cutting-edge technology, then go through the trial-and-error process. Since this trial-and-error process cannot be done alone in many cases, it is necessary to try various directions in collaboration with different people, which requires discussions. I think a solution to a problem can be found by repeating this process of trial and error.

■ **Interviewee profile**

Masaaki Nagata received a B.E., M.E., and Ph.D. in information science from Kyoto University in 1985, 1987, and 1999. He joined NTT in 1987. His research interests include morphological analysis, named entity recognition, parsing, and machine translation. He is a member of the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan, Japanese Society for Artificial Intelligence, the Association for Natural Language Processing, and the Association for Computational Linguistics.