Efforts toward Service Creation with Neural Machine Translation

Sen Yoshida, Masahide Mizushima, and Kimihito Tanaka

Abstract

The accuracy of machine translation has been improving rapidly with the emergence of neural machine translation (NMT) that incorporates deep-learning technology, and it is therefore being utilized in many applications. NTT Media Intelligence Laboratories is working on research and development aimed at creating and enhancing new services using NMT. In this article, we present an overview of NMT and explain some technical points in utilizing it in actual services.

Keywords: machine translation, deep learning, domain adaptation

1. Introduction

Advances in the development and practical application of neural machine translation (NMT), in which deep learning technology is applied, has resulted in substantial improvements in the accuracy of machine translation[*1]. The outputs of translation services that can be readily utilized from web browsers are considered to be significantly better than they were in years past.

Research on machine translation technology has been underway since around the 1950s. In the early years, translation rules were hand-crafted, but statistical machine translation using machine learning has been under development since the 1990s. This has reduced the cost of building translation systems, resulting in the gradual expansion of machine translation services. Furthermore, the emergence and subsequent improvement in translation accuracy of NMT has led to an increase in the range of its practical use. Consequently, machine translation technology is flourishing.

The NTT Group also provides various types of translation services. In addition to the services offered by NTT DOCOMO via mobile phones and other devices [1], Mirai Translate commenced a machine translation service equipped with an NMT engine [2], which is commercially available through NTT EAST's Hikari Cloud cototoba [3] and NTT Communications' AI Translation Platform COTOHA Translator [4].

Implementing NMT is relatively easy because it has a simpler architecture than conventional machine translation, and various open source software (OSS) libraries for deep learning are available today. In recent years at machine-learning-related academic conferences, authors submitting papers were increasingly being subjected to demands to disclose the source code of their programs. This subsequently led to open-source NMT systems such as OpenNMT [5] becoming easily available to anyone.

2. Overview of NMT

We briefly explain here the mechanism of NMT [6] using examples of English-Japanese translation, in which English text is given as input and is output as text translated into Japanese.

To execute an NMT translator, it is necessary to create a translation model in advance. A translation model in NMT is a neural network that converts English text into Japanese. Like other artificial intelligence (AI) technologies such as image recognition and speech recognition, the neural network of NMT is created by learning from a large number of samples.

*1 This article was translated from Japanese using neural machine translation and post-edited by humans.

In NMT, bitexts, that is, pairs of English and Japanese sentences whose meanings are the same, are used as samples. Below is an example of a bitext.

English: A brown dog is running.
Japanese: 茶色い犬が走っている.
(Chairoi inu ga hashitte iru)

We prepare 0.1–100 million of such English-Japanese bitexts and let the neural network program learn them. With a translation model created in this way, a translator program can translate English texts that are not included in the samples into Japanese.

We explain the structure of the NMT neural network using the case of a sequence-to-sequence model with an attention mechanism [7], which is currently a mainstream scheme. In NMT, we first decompose input text into word sequences[*2] as a preprocessing step. Then, each word is mapped onto a vector space of hundreds or thousands of dimensions. Finally, we represent all sorts of information such as the naturalness of the connection between words and the degree of attention given to each word as vectors, and perform translations by doing computations between vectors.

In the process of learning the translation models, the weights in the computation of vectors are adjusted to more appropriate ones. This involves using 0.1–100 million bitexts as learning data. For this reason, an extremely large number of vector operations are performed in the learning of translation models. As with other deep learning technologies, it is therefore common to use a graphics processing unit (GPU)[*3], which is hardware that can perform a large number of vector operations at high speed and can be introduced relatively inexpensively. However, even when GPUs are used, the learning of the translation model takes several days to weeks, which is one of the factors in the high cost of developing translation services using NMT.

3. Technical efforts in service creation

At NTT, we are carrying out research and development (R&D) of NMT as part of efforts to create new services. We describe these efforts in this section.

3.1 Adaptation to a limited target domain

A translation service that can be used easily from a web browser does not specifically limit the domain of texts to be translated, and it covers any text from an academic paper on cancer treatment to a menu in a steak house. However, the accuracy of translation is generally higher by limiting the domain of the text to be targeted to a certain extent, for example, by creating and using a dedicated translation model such as for medical papers or for American tourism guides. For this reason, in developing services using NMT, it is vitally important to determine the target domain of a promising translation-related business and to achieve the required translation accuracy by using a customized translation model, which cannot be substituted by a general-purpose translation model.

The simplest and most powerful way to build a translation model that is customized in a targeted domain is to procure a large volume of bitexts in that domain and use them to create a translation model. NTT Media Intelligence Laboratories has begun efforts to improve translation accuracy by collecting bitexts that exist within the NTT Group or by creating them.

3.2 Fine tuning

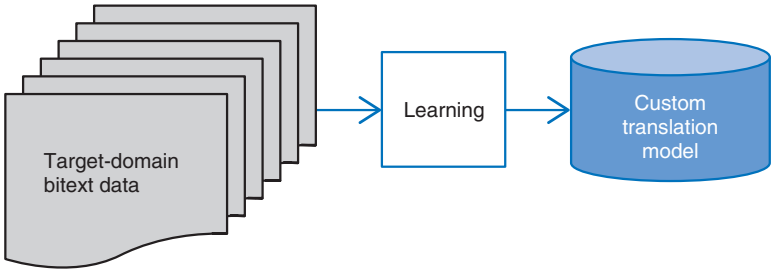
As explained previously, however, building a translation model with bitexts in the target domain from scratch takes several days or weeks. This makes it too costly to adapt to diverse business needs. Therefore, a technique called *fine tuning* is sometimes used as a means of customizing a translation model at low cost by performing further learning using a relatively small quantity (from several thousand to several hundred thousand) of domain-specific bitexts while using an existing translation model as a basis.

Fine tuning is convenient because it makes it easy to build a custom translation model with only a relatively small number of domain-specific bitexts, although at present, it is more likely to be inferior in terms of translation accuracy compared with the case of building a translation model from scratch. This makes it necessary for a translation business using NMT to use the fine tuning and building-from-scratch in a case-by-case manner by considering the amount of time given and the number of bitexts as well as the scope of the target business domain (**Fig. 1**).

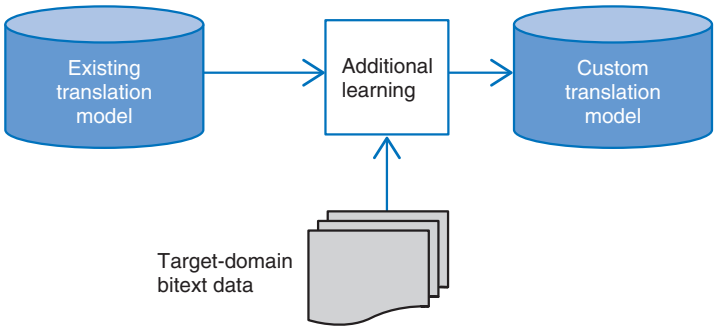
NTT Media Intelligence Laboratories is accumulating know-how on the stabilization of performance in fine tuning, and on the proper use of fine tuning and building-from-scratch.

*2 Word sequence: There are cases in which sentences are segmented into ordinary words and segmented into subwords, which are even finer character strings. Because there is no essential difference in the neural network structure in either case, the segmented character strings are uniformly called "words" in this article.

*3 GPU: Originally an image computing device mounted on a graphics board or the like; however, a technique called general-purpose computing on GPU (GPGPU) that uses GPUs for purposes other than image computation is spreading.



(a) Creating a translation model from scratch



(b) Fine tuning

Fig. 1. Method for creating a custom translation model.

3.3 Improving quality of bitexts

Whether we build a translation model from scratch or perform fine tuning, the nature of bitexts used greatly affects the translation accuracy. The nature of bitexts refers to the degree of word variation and the degree of direct translation.

For the degree of word variation, for example, even a sentence representing the same meaning 犬が走っている (A dog is running) can be written in various ways:

・犬が走っています (Inu ga hashitte imasu)
・イヌが走ってる (Inu ga hashitte ru)

It is generally regarded that the fewer word variations there are in bitexts, especially on the target side (Japanese side in English-Japanese translation), the better. This is because when a neural network learns word sequences and the like, it is easier to learn if the sequences are always similar.

Moreover, with regard to the degree of direct translation of bitexts, the readability for humans is emphasized in normal translation (not for machine-translation learning data), so it is often the case that words that are not actually written in English are complemented from the context and incorporated into the Japanese translations. However, the current form of NMT does not handle contextual information, so such complementation can instead become noise in the learning data. Thus, straightforward direct translation is more suitable as bitext data.

NTT Media Intelligence Laboratories has been building guidelines for ensuring good quality of bitext data by suppressing word variations or increasing the degree of direct translation.

3.4 Making named entities variables

Although NMT learns a translation model from a large amount of bitext data, a problem arises in that some expressions that rarely appear in bitext data cannot be learned well. Such problems become prominent especially with numerical values. Numbers exist infinitely, so no matter how large the amount of learning data we collect, we cannot learn everything from the sample.

Similar problems occur even in person names and place names. These phrases indicating proper nouns (e.g., person names and place names), date and time, and other details are referred to as named entities. In machine translation including NMT, it is common to

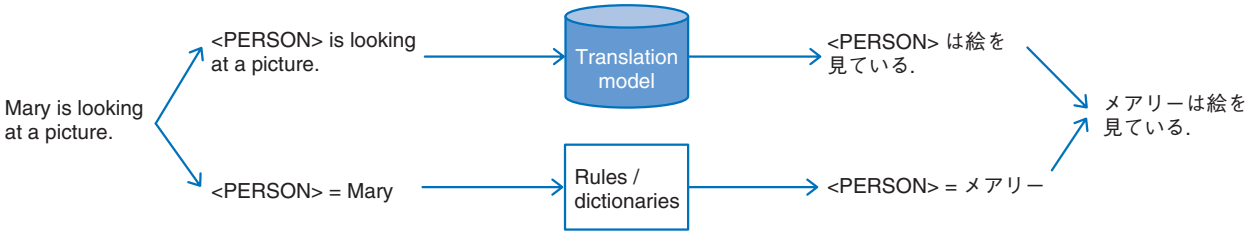


Fig. 2. Converting named entities into variables.

deal with this problem by converting named entities that appear in texts into variables. This is done as follows:

Mary is looking at a picture. →<PERSON> is looking at a picture.

This makes it unnecessary to learn all person names individually in a translation model. However, in a practical translation process, we use a translation model with variables to generate a translation result such as "<PERSON> is looking at a picture," while simultaneously converting "Mary" to "メアリー" and finally returning the variable part to the normal text (**Fig. 2**).

For English-to-Japanese conversion of named entities into variables, for example, a person name and a date, rewriting rules such as regular expression and dictionaries are used. NTT Media Intelligence Laboratories has been working on effective ways of building such rules and dictionaries by utilizing language processing technology such as named entity recognition that we have been developing for many years.

Because treatment of such named entities is not adequately supported by OSS such as OpenNMT, named entity processing is one key to the practical use of NMT. Another important issue is determining how to achieve the functions of a user dictionary, in which users themselves specify phrase translations, in NMT.

4. Future development

NTT Media Intelligence Laboratories will continue to work on R&D aimed at improving the translation accuracy of NMT and adapting it to the business domain efficiently and effectively.

References

[1] Website of translation services by NTT DOCOMO (in Japanese), http://honyaku.idc.nttdocomo.co.jp/
[2] Press release issued by Mirai Translate on December 11, 2017 (in Japanese). https://miraitranslate.com/news/323
[3] Press release issued by NTT EAST on July 3, 2017 (in Japanese). https://www.ntt-east.co.jp/release/detail/20170703_01.html
[4] Press release issued by NTT Communications on January 15, 2018 (in Japanese). http://www.ntt.com/about-us/press-releases/news/article/2018/0115.html
[5] OpenNMT, http://opennmt.net/
[6] Y. Tsuboi, Y. Unno, and J. Suzuki, "Natural Language Processing by Deep Learning," Kodansha, Ltd., 2017 (in Japanese).
[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," Proc. of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, USA, May 2015.

Trademark notes

Sen Yoshida
Senior Research Engineer, Supervisor, Knowledge Media Project, NTT Media Intelligence Laboratories.
He received a B.E. and M. Info. Sci. from Tohoku University, Miyagi, in 1993 and 1995. He joined NTT Communication Science Laboratories in 1995 and studied multi-agent systems and online social systems. During 2003–2007, he worked at NTT WEST as a member of the IPv6-based commercial public network development team. Since 2007, he has been working on R&D of natural language processing systems, including machine translation, at both NTT Communication Science Laboratories and NTT Media Intelligence Laboratories. He is a member of the Information Processing Society of Japan and the Japanese Society for Artificial Intelligence.

Masahide Mizushima
Senior Research Engineer, Knowledge Media Project, NTT Media Intelligence Laboratories.
He joined NTT in 1992. Since 2012, he has been developing natural language processing systems, including machine translation.

Kimihito Tanaka
Senior Research Engineer, Knowledge Media Project, NTT Media Intelligence Laboratories.
He received a B.E. and M.E. from the University of Tsukuba, Ibaraki, in 1993 and 1995. He joined NTT Human Interface Laboratories in 1995 and studied speech synthesis technologies. During 1999–2009, he worked at NTT EAST as a member of the Customer Premises Equipment Business Division and the R&D Center. Since 2012, he has been working on R&D of artificial intelligence systems, including machine translation, at NTT Media Intelligence Laboratories and R&D Planning Department.