Recent Innovations in NTT's Statistical Machine Translation

Masaaki Nagata, Katsuhito Sudoh, Jun Suzuki, Yasuhiro Akiba, Tsutomu Hirao, and Hajime Tsukada

Abstract

English and Japanese have very different word orders, and they are probably one of the most difficult language pairs to translate. We developed a new method of translating English to Japanese that takes advantage of the head-final linguistic nature of Japanese. It first changes the word order in an English sentence into that of a Japanese sentence and then translates the reordered English sentence into Japanese. We found that our method dramatically improved the accuracy of English-to-Japanese translation. We also found that the method is highly effective for Chinese-to-Japanese translation.

Keywords: statistical machine translation, language distance, preordering

1. Introduction

Machine translation is a technology to translate one language into another language by computer. Research on machine translation started in the 1950s, making it almost as old as the computer itself, and many different types of machine translation systems have been developed over the years.

Many web pages nowadays are written in various foreign languages such as Chinese, Korean, and Arabic due to the advancement of Internet technologies. Multinational companies must translate their manuals and product information quickly and accurately into the local languages. It may be said that it is a fundamental desire for human beings to break through language barriers in order to communicate and exchange knowledge. However, previous machine translation systems have not been capable of satisfying the various translations needs of users.

2. From rule-based translation to statistical translation

Previous machine translation systems required the development of large-scale translation rules and bilingual dictionaries for each language pair. This is a labor-intensive task that requires the efforts of dozens of specialists over several years. This kind of machine translation approach is called *rule-based translation*.

It is often said that rule-based translation has reached its limit in accuracy and is difficult to improve further. *Statistical machine translation* is proposed as an alternative to rule-based translation. It automatically learns statistical models, which are equivalent to translation rules and bilingual dictionaries, from a large number of bilingual sentences—on the order of several hundred thousand to several million. It is an emerging technology in which the ultimate goal is to develop a machine translation system for new language pairs or new domains at low cost and in a short period of time. An outline of the statistical machine translation process is shown in **Fig. 1**.

In around 1990, researchers at IBM proposed a machine translation system between French and English using the Canadian Hansard, the transcripts of parliamentary debates. This was the first attempt to use statistical machine translation. In the 2000s, the accuracy of statistical machine translation reached a level of practical use for language pairs with similar word orders, for example French and English, by applying *phrase-based translation*, in which the translation unit changed from words to phrases.

In around 2005, the translation accuracy for language pairs with larger word differences such as

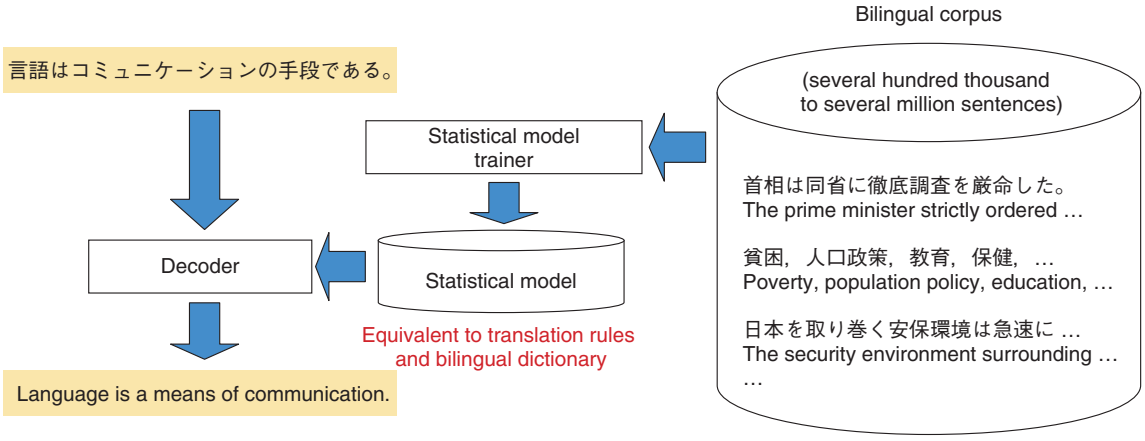

Fig. 1. Outline of statistical machine translation.

Chinese and English were improved by using *tree-based translation*, which uses syntactic theory or the hierarchical structure of either the source or target sentence. The accuracy of statistical machine translation turned out to be higher than that of rule-based translation, not only for language pairs with similar word orders such as French and English, but also for those with relatively different word orders such as Chinese and English. However, statistical machine translation could not outperform rule-based translation for language pairs with highly different word orders such as Japanese and English.

3. Preordering for translation

The idea of *preordering*, in which the word order of the source language sentence is rearranged into that of the target language sentence before translation, was first presented in the early 2000s. It started to gain attention from major research institutes such as Google, Microsoft, IBM, and NTT as a promising technology to overcome word order difference in around 2010. Preordering involves the use of reordering rules in order to obtain the word order of the target sentence. These rules are applied to the syntactic structure obtained by parsing the source language sentence. Reordering rules are usually created manually, although some methods exist for learning them automatically from a bilingual corpus with automatic word alignments.

With respect to reordering, NTT has focused on the head-final nature of the Japanese syntactic structure and has proposed a preordering method for English-to-Japanese translation called head finalization that uses only one rule: *move the syntactic head to the end of the constituent* [1]. We found that it dramatically improves the accuracy of English-to-Japanese translation. In 2011, the joint team of NTT and the University of Tokyo ranked first in the evaluation of an NTCIR-9 (NII (National Institute of Informatics) Testbeds and Community for Information access Research, 9th meeting) patent translation task by combining the University of Tokyo's accurate English parser, Enju, with NTT's preordering technique using head finalization. This was the first time ever that, by human evaluation, the accuracy of statistical machine translation outperformed that of rule-based translation in an English-to-Japanese translation task [2], [3].

4. Preordering method based on head-final order in Japanese

An outline of the preordering method based on the head-final property of Japanese is shown in **Fig. 2**. A phrase is a constituent of a sentence, and its *head* is a word that determines the grammatical role of the phrase in a sentence. For example, a preposition is the head of a prepositional phrase. In other words, in the dependency relation, which all Japanese students learn in Japanese class in elementary school, the word that is modified is the head. In the Japanese pattern of dependency, the dependency always goes from left to right; that is, modified words are always at the sentence-end side of the words that modify them. This is the head-final property of Japanese.

In fact, Japanese is a strictly head-final language, which is very rare in the world. In general, as shown

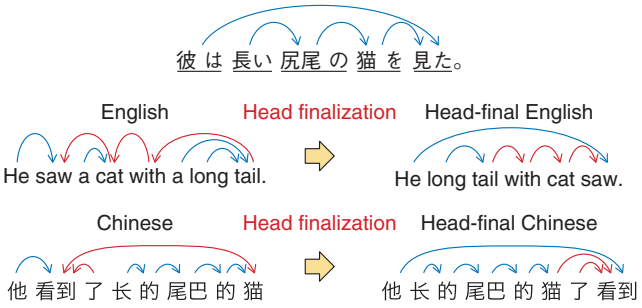

Fig. 2. Preordering method based on head-final property of Japanese.

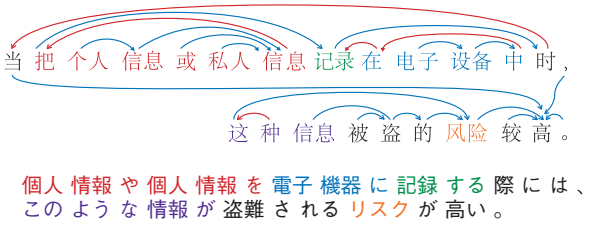

Fig. 3. Example of a dependency structure for a Chinese patent sentence.

in the English and Chinese examples in Fig. 2, dependency goes both from left to right and from right to left. The verb (saw) in the English example is modified by the subject from the left (He) and by the object from the right (cat). With respect to the two nouns, the adjective modifies one of the nouns (tail) from the left, and the prepositional phrase modifies the other noun (cat) from the right. In the Chinese example, with respect to the verb, the subject is on the left and the object on the right, but the modifications of both nouns go from left to right.

Because of this head-final property of Japanese, if we reorder the words of the source language sentence so that its dependency always goes from left to right, the resulting word order is the same as its Japanese translation. This is the basic idea of preordering based on head finalization. If the word order of a source language sentence is the same as that of the target sentence, the remaining task is word-to-word translation, which can be solved accurately by statistical machine translation. Since the head finalization only uses the linguistic properties of the target language, it can be applied to translations from any language into Japanese as long as we have a method to obtain the syntactic structure of the source language.

Translating Japanese into other languages is more difficult than translating other languages into Japanese because for each dependency relation in the syntactic structure of the source Japanese sentence, we have to decide whether to keep its direction or not based on its grammatical role in the target language.

5. Multilingual translation of technical documents

We built a statistical machine translation system to translate patent documents from English, Chinese, and Korean into Japanese in order to verify the feasibility of statistical machine translation from foreign languages into Japanese based on the proposed preordering technique.

In the patent domain, there are *patent families*; this refers to a set of patents for one invention that is applied to different countries that share priority claims to the same patent. Patent documents in a patent family are not a perfect translation of each other, but they include many sentences that are translations of sentences in other documents in the same family. Therefore, we can extract a large-scale bilingual corpus by mining patent families. We prepared three bilingual corpora, English-Japanese (about 4 million sentences), Chinese-Japanese (about 8 million), and Korean-Japanese (about 2 million), from patent documents filed in Japan, the U.S., China, and Korea from 2004 to 2012. As far as we know, the Chinese-Japanese and Korean-Japanese patent corpora are each one of the largest in their language pairs.

To apply the proposed preordering method based on head finalization, we need a technology for parsing the syntactic structure of the source language sentence accurately. We made a set of training data with manually annotated syntactic structures for English (40,000 sentences from news articles and 10,000 sentences from patent documents), and for Chinese (50,000 sentences from news articles and 20,000 sentences from patent documents). We then made a dependency parser for English and Chinese using a semi-supervised learning technique developed by NTT in 2009, which achieved the best published accuracies in international benchmark data for English and Czech dependency parsing [4].

An example of a dependency structure of a Chinese patent sentence is shown in **Fig. 3**, and an example of its Chinese-to-Japanese translation is shown in **Fig. 4**.

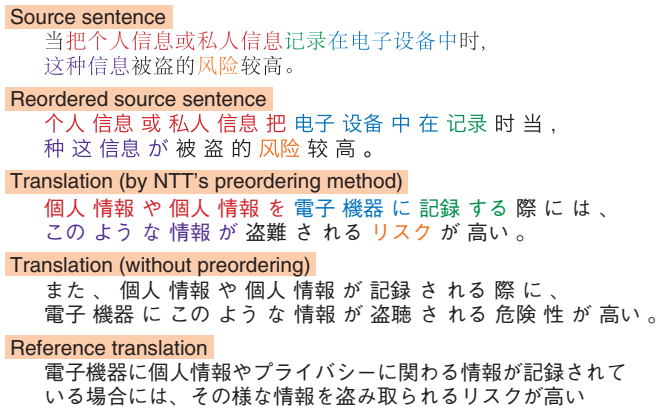

Fig. 4. Example of Chinese-to-Japanese translation of patent sentence.

In general, sentences in patents are long and have complicated dependency structures. However, we found that with head finalization, we could generate a target Japanese sentence that accurately reflected the dependencies in the source Chinese sentence if we could parse the dependency of the Chinese sentence correctly. It should be noted that we did not have to apply preordering in the Korean-to-Japanese translation because the word order in Korean is almost the same as that in Japanese.

6. Automatic evaluation of translation accuracy

Finally, we briefly explain the automatic evaluation of translation accuracy. Objective evaluation of machine translation accuracy is very difficult. There are many correct translations for a sentence, and it is a subjective decision whether to focus on word translation errors or word order errors. An automatic evaluation measure for translation called BLEU (BiLingual Evaluation Understudy), brought a revolutionary change and accelerated the research on machine translation when it was presented in the 1990s. In a sense, the effect was similar to the invention of the instrument for measuring the taste of rice, which activated the competition between rice producing areas and prompted efforts to improve the various breeds. One problem with BLEU, however, is that it does not agree with human evaluations of translations between English and Japanese. The correlation between human evaluation and automatic evaluation by BLEU for the Japanese-to-English translation of a patent translation task in NTCIR-7, which was held in 2008, is shown in **Fig. 5**.

To solve this problem, we proposed a novel automatic evaluation measure called RIBES (Rank-based Intuitive Bilingual Evaluation Score) and released it to the public as open source software [5], [6]. It focuses more on the degree of word order agreement between translation results and reference translations. RIBES was adopted as one of the official evaluation measures at the previously mentioned NTCIR-9, and organizers of the workshop found that it had higher agreement with human evaluation than BLEU in English-to-Japanese, Japanese-to-English, and Chinese-to-English translation tasks [2].

7. Practical application of machine translation

In the translation of technical documents such as patents, manuals, and scientific journals, it is very important to transfer the objective and logical meaning of the content, that is, to accurately map the modifier-modified relations from the source language to the target language. We think that statistical machine translation from foreign languages into Japanese has reached the level of practical use for domains such as patents, in which we can collect a bilingual corpus of more than one million sentences.

For translation from Japanese to English, statistical machine translation is yet to outperform rule-based translation, although the difference in accuracy is getting increasingly smaller. Future tasks include extending the application domains from technical documents to business documents and to spoken languages, as well as improving the translation from Japanese to foreign languages.

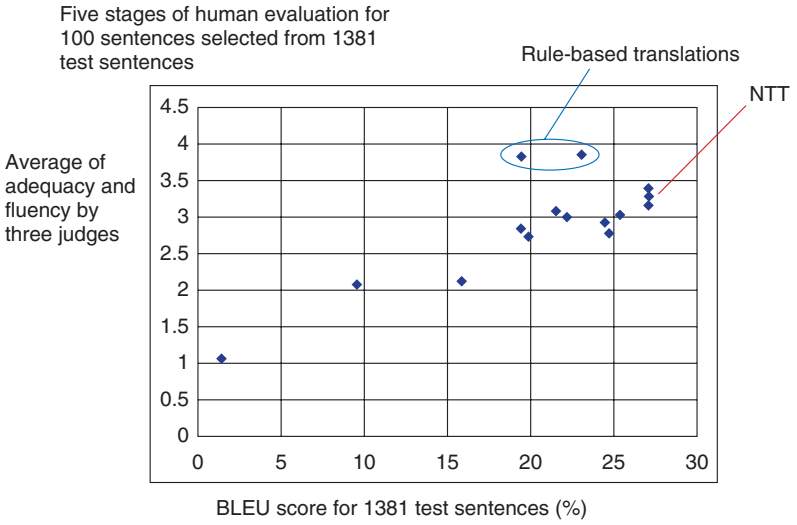

Fig. 5. Correlation between human evaluation and BLEU in NTCIR-7 Japanese-to-English patent translation task (2008).

References

[1] H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh, "HPSG-Based Preprocessing for English-to-Japanese Translation," Journal of ACM Trans. on Asian Language Information Processing (TALIP), Vol. 11, No. 3, 2012.

[2] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou, "Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop," Proc. of NTCIR-9 Workshop Meeting, pp. 559–578, Tokyo, Japan, 2011.

[3] K. Sudoh, K. Duh, H. Tsukada, M. Nagata, X. Wu, T. Matsuzaki, and J. Tsujii, "NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT," Proc. of NTCIR-9 Workshop Meeting, pp. 585–592, Tokyo, Japan, 2011.

[4] J. Suzuki, H. Isozaki, X. Carreras, and M. Collins, "An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing," Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 551–560, Suntec, Singapore.

[5] RIBES: Rank-based Intuitive Bilingual Evaluation Score. http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html

[6] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, "Automatic Evaluation of Translation Quality for Distant Language Pairs," Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 944–952, Cambridge, MA, USA.

Masaaki Nagata

Senior Distinguished Researcher, Group Leader, NTT Communication Science Laboratories.

He received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University in 1985, 1987, and 1999. He joined NTT in 1987. He was with Advance Telecommunications Research Institutes International (ATR) Interpreting Telephony Research Laboratories, Kyoto, from 1989 to 1993. He was a visiting researcher at AT&T Laboratories Research from 1999 to 2000. His research interests include natural language processing, especially morphological analysis, named entity recognition, parsing, and machine translation. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence (JSAI), the Association for Natural Language Processing (ANLP), and the Association for Computational Linguistics (ACL).

Katsuhito Sudoh

Research Scientist, NTT Communication Science Laboratories.

He received the B.Eng. and M.Inf. degrees from Kyoto University in 2000 and 2002. He joined NTT in 2002 and studied spoken dialogue systems and spoken language processing. He is currently working on statistical machine translation. He is a member of IPSJ, ANLP, ACL, and the Acoustic Society of Japan (ASJ).

Jun Suzuki

Senior Research Scientist, NTT Communication Science Laboratories.

He received the B.Sc. degree in mathematics and M.Eng. degree in computer science from Keio University, Kanagawa, in 1999 and 2001 and the Ph.D. degree in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology in 2005. He joined NTT Communication Science Laboratories in 2001. He is currently studying machine learning and natural language processing areas including kernel methods, supervised and semi-supervised learning, question answering, machine translation, and natural language parsing. During 2008–2009, he stayed at MIT CSAIL in Boston as a visiting researcher to develop the high-performance dependency parser. Since September 2013, he has been a member of the editorial board of ANLP Journal. He is a member of IPSJ, ANLP, and ACL.

Yasuhiro Akiba

Senior Research Scientist at NTT Communication Science Laboratories and Senior Research Engineer at NTT Media Intelligence Laboratories.

He received the B.Sc. and M.Sc. degrees in mathematics from Waseda University, Tokyo, in 1988 and 1990 and the Ph.D. degree in informatics from Kyoto University in 2005. He joined NTT in 1990. He was with the ATR Spoken Language Translation Research Laboratories as a Senior Researcher from October 2000 to March 2005. His research interests include machine learning, knowledge acquisition, natural language learning, machine translation, and automatic evaluation. He received the Best Paper Award at the 9th Annual Conference of the Japan Society for Artificial Intelligence in 1995 and the CV Ramamoorthy Best Paper Award of the 12th IEEE International Conference on Tools with Artificial Intelligence in 2000. From April 1999 to March 2001, he was a member of the editorial board of IPSJ Magazine. He is a member of IPSJ and ANLP.

Tsutomu Hirao

Research Scientist, NTT Communication Science Laboratories.

He received the B.E. degree from Kansai University, Osaka, in 1995, and the M.E. and Ph.D. degrees in engineering from Nara Institute of Science and Technology in 1997 and 2002. He joined NTT Communication Science Laboratories in 2000. His current research interests include Natural Language Processing and Machine Learning. He is a member of IPSJ, ANLP, and ACL.

Hajime Tsukada

Senior Research Scientist, NTT Communication Science Laboratories.

He received the B.S. and M.S. degrees in information science from Tokyo Institute of Technology in 1987 and 1989. He joined NTT Human Interface Laboratories in 1989. In 1997, he joined ATR Interpreting Telecommunications Research Laboratories, and from 1998 to 1999, he was a visiting researcher at AT&T Laboratories Research. Since 2003, he has been with NTT Communication Science Laboratories. His research interests include statistical machine translation as well as speech and language processing. He is a member of IEICE, JSAI, ANLP, ACL, and ASJ.