Machine Translation | Spoken Language Translation | Preprocessing Technology

Technology Reports

Spoken Language Translation Services for Diverse Usage Scenes

The number of foreign visitors to Japan continues to increase with expectations of 40 million visitors by 2020. With such a large number of foreign guests coming to Japan, there are hopes for an environment that can provide stress-free communication. At the same time, the globalization of business is accelerating resulting in even more situations where multilingual communication is needed.

Among the translation services that NTT DOCOMO has decided to develop, this article describes the issues and solutions surrounding meeting translation, SNS translation, and customer-reception translation as services that translate spoken language for diverse usage scenes.

Service Innovation Department
Masato Takeichi
Takaya Ono
Yuki Chijiwa
Yixin Jiang

1. Introduction

The number of foreign visitors to Japan in 2015 reached 19,740,000, which surpassed the previous record set the year before in 2014 [1]. With this in mind, the Japanese government has set 40 million visitors as the target for 2020 with expectations that an environment enabling all foreign visitors to communicate freely can be provided [2]. At the same time, the number of employees of overseas subsidiaries of Japanese firms increased to 5,750,000 in FY2014 from 4,990,000 in FY2010 [3], which reflects the onward march of corporate globalization in which multilingual communication will be increasingly needed.

With a view to 2020, NTT DOCOMO has undertaken the development of speech recognition*1 technology and machine translation*2 technology and services that apply those technologies toward means of communication that can surmount language and cultural barriers.

In **Figure 1**, translation services are classified along two axes representing "spoken language – written language" and "hard/soft textual styles of writing" to clarify the technical issues that must be addressed. NTT DOCOMO sets the upper-right area of the figure—"spoken language" and "soft"—as a near-term target with a focus on foreign visitors to Japan and aims to improve the accuracy of speech recognition and machine translation to meet this target. The following three services are currently being developed as part of this endeavor.

(1) Meeting translation: Targeting meetings in different languages, this service translates meeting speech into the user's native language for verbal readout and

*1 **Speech recognition:** Technology for converting speech signals generated by human utterances into text.

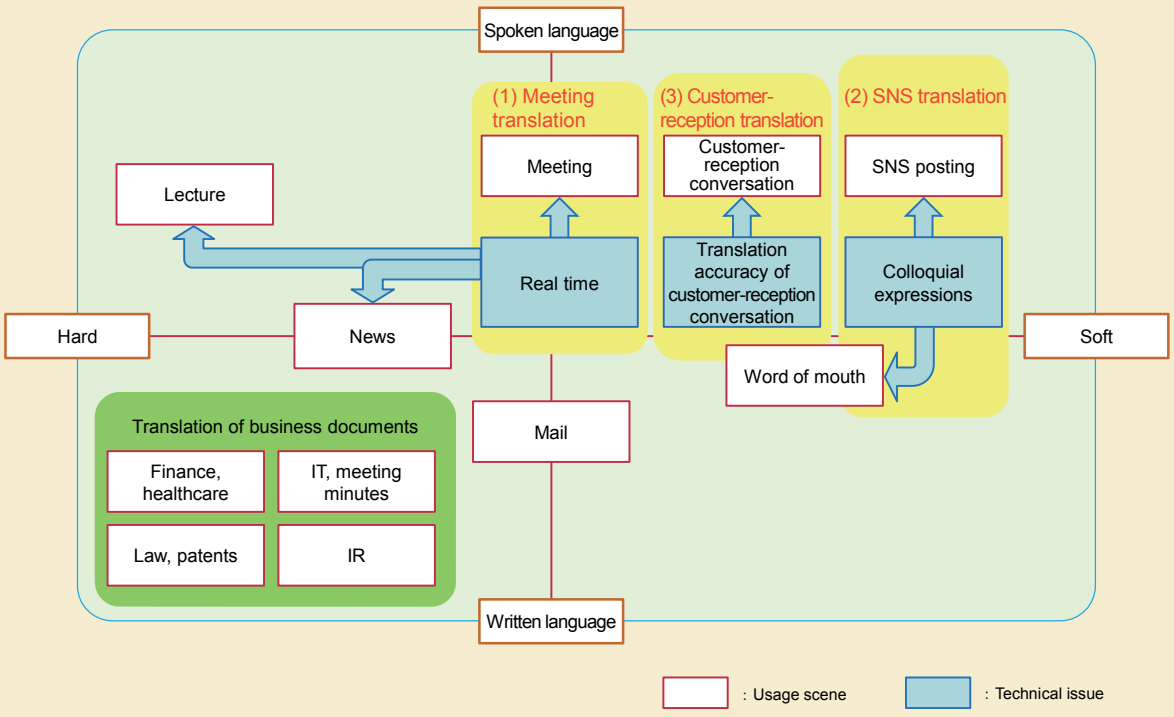


Figure 1 Usage scenes of translation services and technical issues

text display.

(2) SNS translation: This service provides a text translation of SNS posts that are normally full of colloquial expressions. Although it may appear at first glance that such posts consist of written language, they actually include many expressions that can be classified as "soft" or "spoken language."

(3) Customer-reception translation: This service performs speech translation of customer-reception conversation between a customer and staff at a store or other business site. For simple communication needs, it provides an easy means of waiting on a customer through speech recognition and machine translation, and for complex conversation needs, it provides a means of talking with a remote interpreter. Both of these means combined enable accurate and stress-free communication. Since the development of a prototype in 2014, NTT DOCOMO has performed trials with a number of companies and has repeatedly improved the User Interface (UI)*3 and raised the accuracy of speech recognition and machine translation. A commercial service was finally launched in June 2016 as "Hanashite Hon'yaku for Biz"*4.

In this article, we provide an overview of these three services, introduce issues associated with each of these services, and describe NTT DOCOMO's approach to resolving those issues.

2. Overview of Each Translation Service and Associated Issues

2.1 Overview of Translation Services

A translation service normally combines three technologies: speech recognition, machine translation, and speech

*2 **Machine translation:** Technology for mechanically converting sentences and words in a certain language into another language and outputting the results. There are two main methods: rule-based machine translation and statistical machine translation.

*3 **UI:** Operation screen and operation method for exchanging information between the user and computer.

*4 **Hanashite Hon'yaku for Biz:** An NTT DOCOMO business-oriented translation service supporting customer reception by Japanese store clerks for foreign visitors to Japan. Launched in June 2016. A registered trademark of NTT DOCOMO, INC.

synthesis*5. The system configuration of such a service is shown in **Figure 2**.

(1) Speech recognition technology is used to convert speech to text. This process first identifies an utterance by utterance-interval detection*6 where the end of an utterance is established by a silent interval*7 lasting longer than a certain period of time. It then removes noise in the background by noise removal*8 technology. Finally, it converts the results of speech recognition to text using an acoustic model*9 and language model*10 in the speech recognition engine.

(2) Machine translation technology is used to mechanically translate input text into text in another language. This process begins with a preprocessing step that performs syntactic analysis*11 and named entity classification*12 against the pre-translated text to improve translation accuracy. It then uses a machine translation engine*13 and its translation model*14 and language model to translate the input text into text of the designated language. Finally, as a post-processing step, it uses bilingual data from a dictionary to perform substitution of previously classified words before preparing and outputting the results of translation.

(3) Speech synthesis technology produces artificial speech data from input text (results of translation) and verbally reads out that data.

The total corpus of a translation service consists of a corpus for speech recognition and a bilingual corpus for machine translation. The former consists of sets of uttered speech and text created by logs accumulated during service provision. The latter consists of bilingual sets of text in the pre-translation language and text in the post-translation language in units of sentences (e.g., “いらっしゃいませ” ⇔ May I help you?).

To improve the accuracy of speech

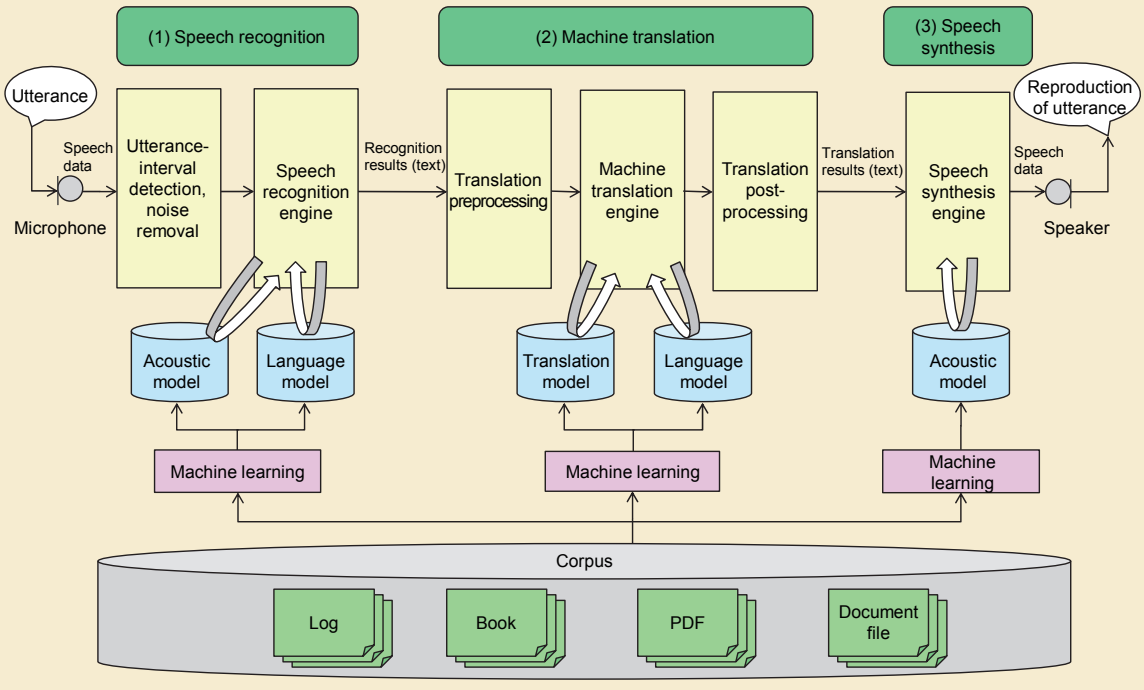


Figure 2 System configuration of speech translation technologies

*5 **Speech synthesis:** Technology for artificially creating speech data from text and verbally reading out text.

*6 **Utterance-interval detection:** Technology for determining the intervals in speech signals in which speech exists and those in which it does not exist.

*7 **Silent interval:** Interval determined to be absent of speech.

*8 **Noise removal:** Process of removing street noise, utterances of other people, etc. to recognize the speech of a specific person.

*9 **Acoustic model:** Statistical model comprising frequency characteristics of phonemes targeted for recognition possesses.

*10 **Language model:** Statistical model comprising morpheme arrangements and frequency of those arrangements.

*11 **Syntactic analysis:** Technology for analyzing the structure of inter-phrase dependency and clarifying grammatical relationships in a sentence.

recognition and machine translation, it is generally necessary to collect a large volume of sentences for each language pair*15 (e.g., Japanese/English, Japanese/Chinese, Japanese/Korean, etc.) as a bilingual corpus of frequently used sentences in usage scenes marked by the red frames in Fig. 1 (such as meetings, SNS posts, and customer reception). Such a large bilingual corpus can then be used as training data in machine learning*16 so that translation models and language models can be created specific to usage scenes. Likewise, in speech recognition, machine learning can be used to create acoustic models and language models using corpuses corresponding to different usage scenes as learning data. In speech synthesis, too, machine learning can be used to create acoustic models.

2.2 Issues Associated with Each Translation Service

The issues particular to each translation service are described below.

(1) Meeting translation must be able to display the results of speech recognition and machine translation in real time and correctly detect the breaks between utterances. With prior technology, the results of applying speech recognition to the utterances of a meeting participant would be displayed in text after an utterance was completed, which would give the user a strong sense of having to wait until results were displayed.

(2) A major issue in SNS translation is dealing with colloquial expressions that frequently appear in SNS posts. Prior technology was incapable of dissecting pre-translated Japanese text written in a non-standard, colloquial format such as "おたんじょーびおめでとぉー" and "これはヤヴァい."

(3) A customer-reception translation service must achieve courteous interaction with customers by store staff and deal with colloquial questions from customers. However, there is a lack of bilingual corpuses for conversation that occurs when waiting on customers, so the issue here is finding ways of improving the accuracy of machine translation for dialog spoken in customer-reception scenarios.

3. Meeting Translation

As shown in **Figure 3**, the meeting translation service performs speech recognition in real time of the conversation of meeting participants speaking different languages and translates that conversation into each other's language. It is also capable of text translation by keyboard input so that a user can participate in a meeting by text if speaking out loud is not possible for whatever reason.

1) Improvement of Real-time Characteristics

To achieve a meeting with good tempo, it is desirable that the processing delay experienced by meeting participants for both speech recognition and machine translation be made small to provide good real-time characteristics.

For this reason, the speech recognition process sequentially displays the results of maximum likelihood estimation for each morpheme*17. This has the effect of minimizing the delay between the utterances of meeting participants and the display of those utterances thereby improving the real-time characteristics experienced by the participants.

2) Automatic Detection of Utterance Interval

Some services that employ speech recognition have the user press a button to mark the beginning and end of an utterance to unambiguously indicate the user's utterance interval. However, for a meeting translation service in which utterances are continuously and rapidly exchanged among meeting participants, a more desirable approach is to automatically detect a speaker's utterance interval and forgo the use of start/stop buttons.

In addition, utterances may be delimited by the continuance of a silent interval that occurs while the speaker is thinking about the next thing to say or

*12 **Named entity classification:** Technology for replacing named entities in the input sentence with labels expressing proper nouns before machine translation and for replacing those labels using a dictionary after machine translation.

*13 **Machine translation engine:** Software for statistically translating text using a language model and translation model trained for machine translation.

*14 **Translation model:** Statistical model used for calculating the extent to which words in a pair of sentences in the pre-translation language and post-translation language semantically correspond with each other.

*15 **Language pair:** A combination of two languages as the translation source and translation objective (e.g. English/Japanese).

*16 **Machine learning:** A mechanism allowing a computer to learn the relationship between inputs and outputs, through statistical processing of example data.

*17 **Morpheme:** Smallest meaningful unit of a language.

simply hesitating. Specifically, in the case of **Figure 4** (1), the Japanese spoken in this example was divided into two utterances consisting of "このアプリは音声認識の結果を" and "リアルタイムに表示します." As a consequence, the machine translation also output two utterances consisting of "This application is speech recognition results." and "Real-time display." In other words, the results output by machine translation in this case differed from the speaker's intention.

To solve this problem, we adjusted the parameter that judges the silent interval so that input speech could be detected as a continuous utterance even with the occurrence of some hesitation

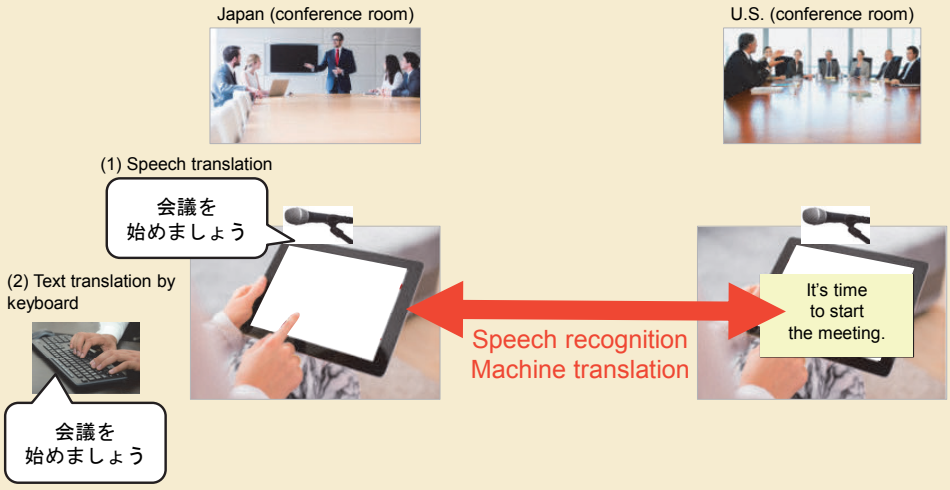
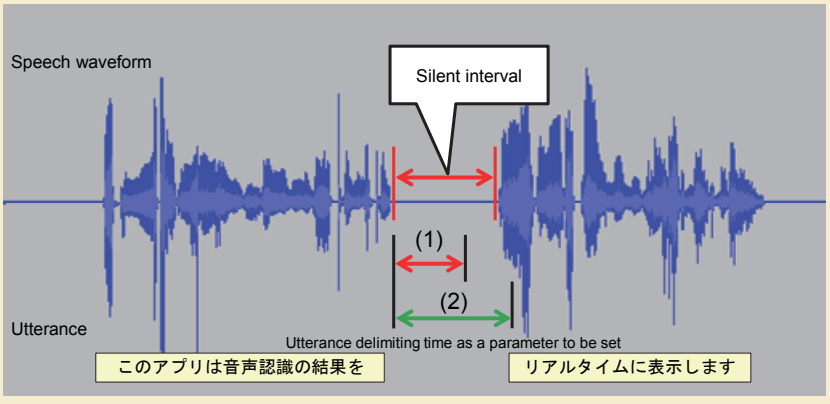


Figure 3 Overview of meeting translation service



Translation results

Case (1)
「このアプリは音声認識の結果を」→ "This application is speech recognition results."
「リアルタイムに表示します」→ "Real-time display."
Case (2)
「このアプリは音声認識の結果をリアルタイムに表示します」
→ "This application is displayed in real-time voice recognition results."

Figure 4 Silent interval caused by hesitation

while speaking. We did this by extracting the silent intervals that occur when speakers hesitate while speaking in meetings and calculating their average value. We then used this average value as a basis for optimizing the parameter to be set as the time for delimiting utterances and performed a test to see if utterances in a meeting could be correctly delimited.

As shown in Fig. 4 (2), adjusting the parameter in this way prevented the silent interval from being treated as an utterance delimiter resulting in the single Japanese utterance "このアプリは音声認識の結果をリアルタイムに表示します." The correct machine translation could then be output as "This application is displayed in real-time voice recognition results."

4. SNS Translation

SNS translation is a service that can translate colloquial sentences and expressions in social media to other languages. The process flow of the SNS translation system is shown in **Figure 5**. This system differs from the other translation services introduced here in that it consists only of the machine translation section shown in Fig. 2.

However, the system adds two functions to translation preprocessing in the usual machine translation section. The first one is sentence dissection and the second one is normalization of variant character strings.

- In addition to identifying ordinary sentence delimiters, the sentence dissection function determines the presence of parentheses, emoticons (e.g. ヽ(・∀・)ﾉ キャッ), service-unique symbols (e.g. RT), URLs, and onomatopoeic/mimetic words in a sentence. As a result, only that part of the sentence recognized as text for translation can now be passed to the following normalization function for processing.
- The normalization function enhances the morphological analysis of traditionally written, standard Japanese by dividing the sentence into "tokens" using variant-token morphological analysis [4] [5] that can deal with colloquial expressions. Here, "tokens" are the result of dividing the input text into minimal word units of the

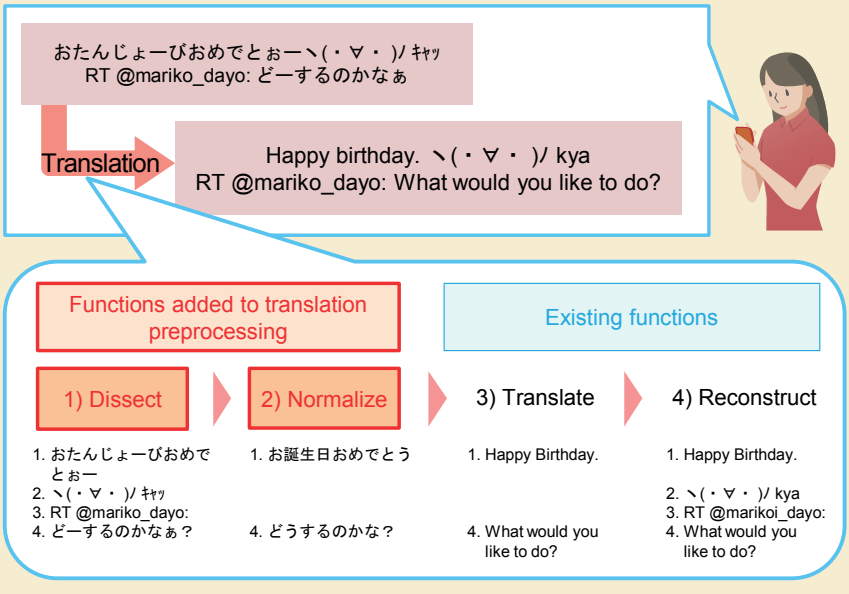


Figure 5 Process flow of SNS translation system

Japanese language. Next, the function generates a representative-token selection lattice*18 using a token variation dictionary and conversion-candidate control list. It then converts the variant token into an optimal token using a language model based on the huge corpus used in machine learning by the machine translation engine. This function can also deal with ambiguity by substitution processing based on negative/positive determination.

1) Generation of Representative-token Selection Lattice and Conversion of Variant Tokens

We here explain the mechanism of generating a representative-token selection lattice and converting variant tokens using **Figure 6**.

First, referring to Fig. 6 (1) and (2), the results of variant-token morphological analysis are used to prepare tabular data consisting of token, Parts Of Speech (POS), and standard token*19 entries from the input sentence. In this example, the input sentence "おたんじょーびおめでとぉー" is divided into the tokens "お," "たんじょーび," "おめでとぉ," and "ー," and the standard tokens "御," "誕生日," "おめでとう," and "ー" are given for each of these. Here, "standard token" refers to standard notation as used in newspapers and other publications.

Next, focusing on standard tokens, a lattice of selectable tokens is created using tokens in the conversion-candidate control list. This list includes candidates not desirable for conversion and candidates uniquely desirable for conversion

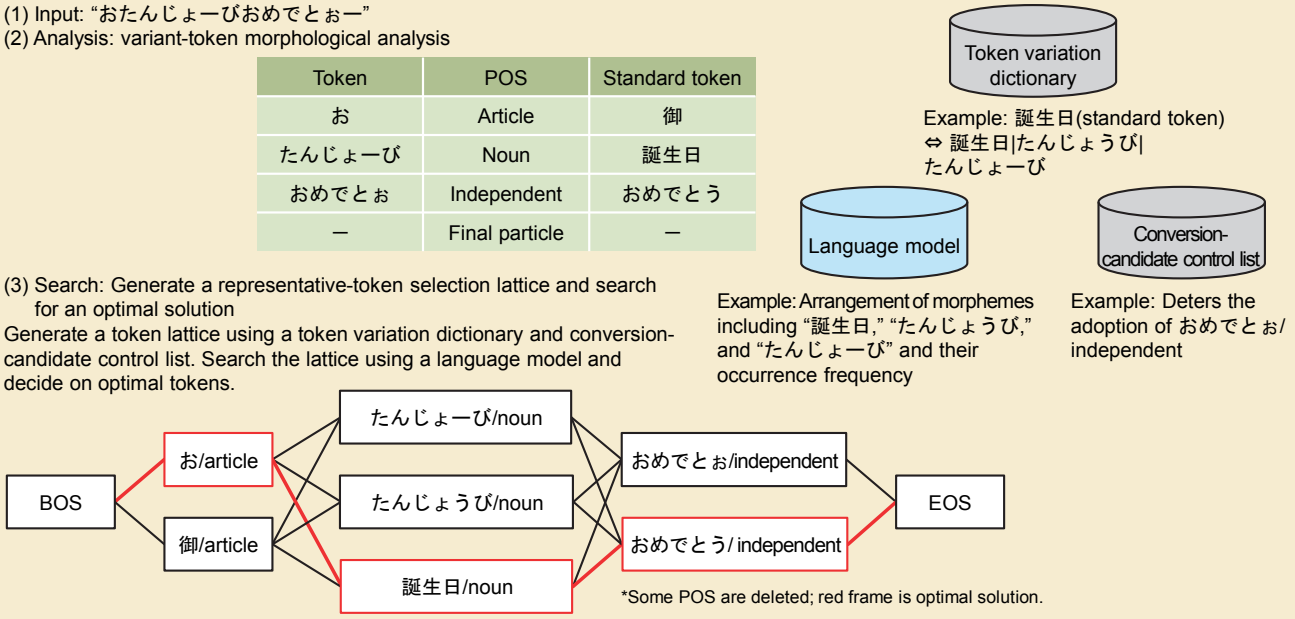


Figure 6 Mechanism of variant-token morphological analysis and variant-token conversion

*18 **Lattice:** A lattice-shaped collection of data that arranges a series of morphemes in the horizontal direction and morphemes with the same meaning but different notation in the vertical direction.

*19 **Standard token:** A characteristic string expressed in characters and symbols that are written in standard format according to undeviating grammar.

as tokens linked to the standard token but not included in the input sentence. For example, tokens linked to the standard token "誕生日" include "たんじょーび," "たんじょうび," and "誕生日," all of which are used to create the token lattice shown in Fig. 6. Here, "token lattice" refers to a graphical structure that enumerates selectable tokens. Now, using a language model employed by the machine translation engine, the strings of tokens in this token lattice are searched to determine an optimal sequence of tokens, as shown in Fig. 6 (3). In this example, the optimal solution is found to be the tokens "お," "誕生日," and "おめでとう," as shown in Fig. 6 (4). This representative-token conversion process results in the normalization of variant character strings.

Incidentally, based on technical support and process results provided by NTT Media Intelligence Laboratories, we have incorporated the language models and conversion-candidate control list created by NTT DOCOMO in this variant-token morphological analysis function and representative-token conversion function and have implemented these functions in the translation preprocessing section.

2) Negative/positive Determination

Japanese can sometimes be ambiguous making it difficult to determine intention from one word or even one sentence. For example, a SNS post stating "これはヤヴァい (Kore wa yabai!)" in relation to a food review can be interpreted as either "delicious" or "bad tasting" if translating one sentence at a time. Accordingly, by applying negative/positive determination*20 to previous and subsequent sentences, the above expression can be substituted by "This is great!" for a positive sentiment and "This is terrible!" for a negative sentiment. Such substitution processing can deal effectively with ambiguity.

In SNS translation, machine translation is applied only to those portions of the input sentence that have been normalized by the above sentence dissection function and variant-character-string normalization function. In translation post-processing, the structure of the input sentence is reconstructed integrating the results obtained in the machine translation step and the reconstructed sentence is returned to the user as the translated sentence.

As a result of this translation post-processing, the accuracy of translating "soft" spoken language from Japanese to English, Chinese, and Korean was found to outperform typical engines of other companies. The results of comparing the accuracy of Japanese/English translation between NTT DOCOMO's SNS translation engine and another company's engine are shown in **Figure 7**. For this comparison, we extracted 100 sentences randomly from SNS posts related to food, cosmetics, and travel and performed a subjective evaluation using three subjects based on the following criteria: "○ (3 points): correctly translated" and "△ (2 points): some errors noticed but meaning is conveyed." The average of the scores given by the three subjects was calculated for each sentence. On tabulating the total score for all sentences, it was found that NTT DOCOMO's SNS translation engine outperformed the other company's engine by 45 points.

5. Customer-reception Translation

Customer-reception translation requires high accuracy in machine translation for both the case of translating courteous sentences spoken by a Japanese-speaking clerk into multiple languages and the case of translating questions that may not necessarily be formal or courteous in the customer's foreign language into Japanese. With this in mind, we created a translation model especially for customer reception by collecting sentences frequently spoken when waiting on customers as a bilingual corpus and having a translation engine learn those sentences. We also applied this bilingual corpus to the language model of the speech recognition engine so that it too could deal with conversation in customer reception.

The following describes the process flow for compiling the bilingual corpus.

*20 **Negative/positive determination:** Method for determining whether the writer's intention is negative or positive from that document.

1) Collection of Common Bilingual Corpuses

The flow of collecting commonly used bilingual corpuses is shown in **Figure 8**. The first step in this process is to collect bilingual data as the basis of a bilingual corpus by the following methods:

- Transcribe and collect logs of commercial services (Hanashite Hon'yaku, mail translation, etc.) in conformance with terms of service
- Purchase and use external bilingual corpuses

The next step is to format the bilingual data collected by these two methods

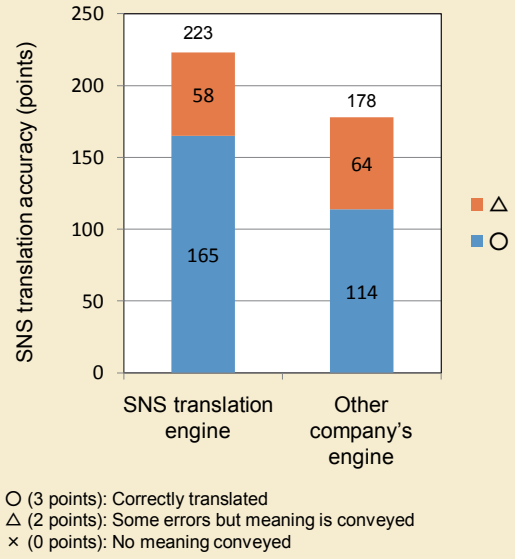


Figure 7 Subjective evaluation of SNS translation accuracy (J-E translation)

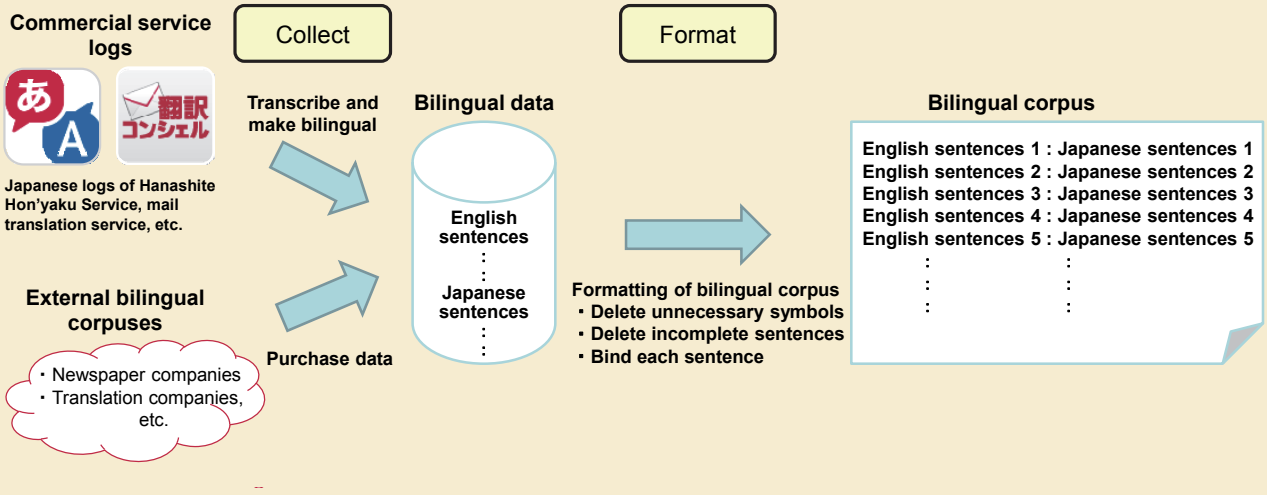


Figure 8 Compilation of bilingual corpus

in the following ways to create a bilingual corpus:

- Delete unnecessary symbols included in the collected corpuses
- Delete incomplete sentences
- Bind each sentence

2) Create a Bilingual Corpus Manually

This time, with the aim of increasing corpus content specialized for a certain usage scene in addition to using the usual collection methods described above, we employed crowdsourcing*21 and other techniques to manually create a bilingual corpus assuming customer-reception conversation. Using the bilingual corpus obtained in this way, we created a translation model and a speech-recognition language model for use in customer reception.

As a result, the accuracy of speech recognition and machine translation was improved for a variety of languages. The results of comparing the accuracy of Japanese/English translation among various engines are shown in **Figure 9**. For this comparison, we randomly extracted 200 sentences from documents and logs related to customer reception and performed a subjected evaluation using five subjects. As shown, the speech-recognition/machine-translation engine that we prepared for customer reception achieved an average value of 218.4 points, which outperformed typical engines of other companies.

6. Conclusion

In this article, we described NTT DOCOMO's approach to achieving spoken language translation and solving associated technical issues with a focus on meeting translation, SNS translation, and customer-reception translation.

Going forward, we plan to refine this technology to make it even easier to use. We also plan to research and develop technologies for improving translation accuracy even further. These will include "prereordering technology" for reordering the words in the target sen-

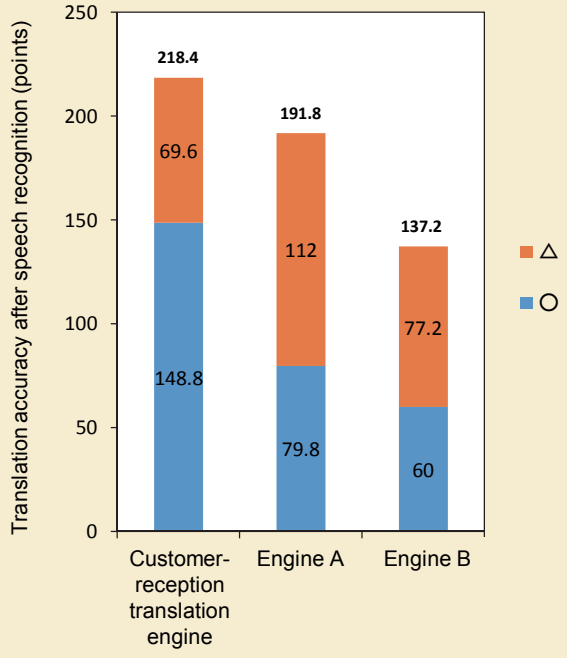


Figure 9 Subjective evaluation of customer-reception translation accuracy (J-E translation)

*21 **Crowdsourcing:** A coined term combining "crowd" and "outsourcing" referring to a new employment format that consigns work or tasks among a number of widely distributed people.

tence prior to translation to mitigate drops in accuracy caused by differences in word order between the source language and destination language, and natural language processing technology for filling in abbreviations and omissions that frequently occur in spoken language.

References

[1] Japan Tourism Agency: "White Paper on Tourism in Japan (2016)."

[2] Prime Minister of Japan and His Cabinet: "Meeting of the Council for the Development of a Tourism Vision to Support the Future of Japan," Mar. 2016. http://japan.kantei.go.jp/97_abe/actions/201603/30article1.html

[3] Ministry of Economy, Trade and Industry, Minister's Secretariat, Research and Statistics Department, Structural Statistics Office, Trade and Economic Cooperation Bureau, Trade and Investment Facilitation Division: "Summary of the 45th Basic Survey on Overseas Business Activities—FY2014 Results—," Jul. 2015. http://www.meti.go.jp/english/statistics/tyo/kaigaizi/pdf/h2c408je.pdf

[4] I. Saito, K. Sadamitsu, H. Asano and Y. Matsuo: "Japanese Morphological Analysis Based on Alignment of Standard and Variant Character Strings and Variant Token Normalization using Character-type Conversion," Proc. of the 20th Annual Meeting of The Association for Natural Language Processing, pp.777-780, Mar. 2014 (in Japanese).

[5] I. Saito, K. Sadamitsu, H. Asano and Y. Matsuo: "Token Normalization and Morphological Analysis using Occurrence Probability of Variant Tokens," Proc. of the 21th Annual Meeting of The Association for Natural Language Processing, pp.51-54, Mar. 2015 (in Japanese).