# An Automatic Interpretation System for Travel Conversation

*Takao Watanabe, Akitoshi Okumura, Shinsuke Sakai,*

*Kiyoshi Yamabana, Shinichi Doi, Ken Hanazawa*

NEC Corporation

## ABSTRACT

We have developed an automatic interpretation system running on a mobile PC that helps oral communication between Japanese and English speakers in various situations during their travel abroad. In order to allow a wide range of expressions and topics in the applied domain, we adopted an approach which utilizes the general linguistic knowledge as well as the domain-specific linguistic knowledge. Speech recognition module performs speaker-independent large-vocabulary (50,000 Japanese words and 10,000 English words) continuous speech recognition. Translation module performs syntax directed translation based on a new lexicalized grammar rule formalism.

## 1. INTRODUCTION

Automatic interpretation is one of prospective applications of the speech and language technology and there are several works on this topic[1,2,3]. Recently speech recognition technology has started to be applied to various practical applications. However, in order to develop a practical automatic interpretation system which can help oral communication between speakers of different languages during their travel abroad, the problem of the limitation in the variety of acceptable situations and expressions needs to be solved. We have tried to develop an automatic interpretation system with an approach to allow a wide range of expressions and topics in the application domain achieving high quality of the interpretation.

For a relatively small-scale domain, an approach which uses domain-specific concept(inter-lingua) such as ‘speech act’ or ‘semantic frame’ is effective[4]. It is relatively easy to obtain such a conceptual expression for an input utterance by applying speech understanding techniques and to generate a sentence in a target language for the domain. However, this approach would not be applicable to a larger scale of domain because the design of such domain-specific concept which is rather difficult.

On the other hand, speech recognition technology for dictating text and machine translation technology for document translation are widely available today, which are basically designed for general-purpose, that is, domain–independent use. One approach to deal with a large-scale domain is to combine these technologies through the interface expressed as a text form, where a speech recognition system and a translation system deal with conversational speech and domain specific expressions.

One idea for this is to derive the knowledge for the speech recognition system and the translation system from a large amount of domain knowledge base, in particular “domain corpus”. This idea should be hopeful if a sufficiently large amount of corpus is collected. However the amount of the corpus which can be collected is inevitably insufficient for a large-scale domain.

We adopted an approach which utilizes the general linguistic knowledge as well as the domain-specific linguistic knowledge. In speech recognition, on the basis of statistical language modeling, we developed a language model using both the domain knowledge, that is domain corpora, and general linguistic knowledge, and incorporate it into large vocabulary continuous speech recognition. In translation, we developed a new lexicalized grammar formalism, which is suitable to handle pattern-like expressions specific to the domain conversations as well as general expressions. We adopted direct language-pair based translation approach rather than inter-lingua approach.

Compact implementation was another issue to be concerned, since it is essentially important during a travel use. With these in mind, we have developed the interpretation system as a compact software on a mobile PC.

## 2. SYSTEM OVERVIEW

### 2.1. Functions

The system was developed to assist travelers with communication in a wide variety of situations. Users can have their speech simultaneously translated in real-time by a mobile PC from either Japanese or English to the other. Together with an 50,000 Japanese and 10,000 English word vocabulary, the software allows users to speak naturally without restriction.

To reduce misunderstanding in the conversation between users talking with each other through the system and to avoid the halt in the conversation, the system accepts input of any utterance unit other than a sentence, namely a fragment of a sentence, a phrase, or a word. For each utterance, translated result is obtained in real-time. Users can confirm the recognized result. Users can also request the system to re-translate after editing the recognized results utilizing the functions such as deleting a portion of the recognized result or inserting text by an additional utterance.

### 2.2. System Configuration

The system consists of four modules: speech recognition, translation, speech synthesis and system integration as is shown in Figure 1. Recognized or translated results are passed between

modules in the form of a text with the supplemental information such as pausing.

Japanese speech synthesis module uses a text-to-speech conversion software developed in NEC. English speech synthesis module uses a commercially available software.

The system integration module consists of three functions: user interface control, inter-module data communications and interpretation module control. Users specify their gender for speech recognition. The users can also specify their role in the situation. Integrated processing for control by situation provides more accurate interpretation.

System requirements for the software include a mobile PC with a Pentium II-class processor (400MHz) running either Windows 98 or Windows NT and 192MB of RAM.

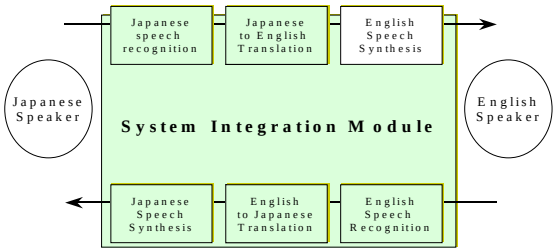



Figure.1 Configuration of the automatic interpretation system

# 3. SPEECH RECOGNITION

## 3.1. Speech Recognition Module

Speech recognition module performs speaker-independent large-vocabulary continuous speech recognition of conversational Japanese and English. The module, shown in Figure 2, consists of an acoustic model, a language model, a word dictionary and a search engine. The acoustic model used was designed domain independently. The language model was designed for travel conversation.

We have developed a compact recognition module that enables real-time speech recognition on a mobile PC. The search engine performs two-stage processing. The language model contains a bigram language model and a trigram language model. On the first stage, Viterbi beam search is performed to decode input speech to generate a word candidate graph using the acoustic model and the bigram language model. On the second stage, the engine performs a search to find the optimal word sequence using the trigram language model.

For acoustic modeling, triphone-context phone HMM was adopted. We have made possible speaker-independent recognition by training the model with a large speech corpus. The recognition module also has a speaker adaptation capability. It is possible to adapt the acoustic models efficiently to the speaker just using as few as five utterances.

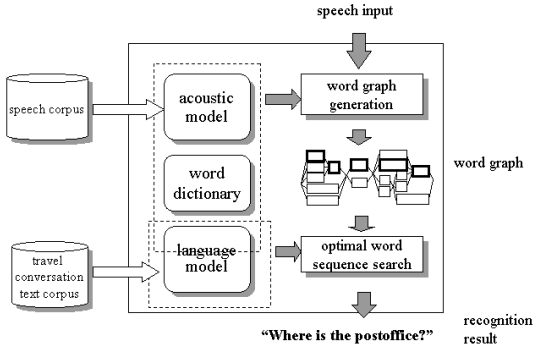



Figure 2. Speech recognition module

## 3.2. Language Model

Statistical language model such as bigram and trigram has been successfully used to give linguistic constraints in large-vocabulary continuous speech recognition and trained typically using a large text corpus such as newspaper. However, colloquial expressions often used in travel conversation, including Japanese sentence-final phrases, are different from written sentences in many ways. Considering this difference, we developed large linguistic corpora. The corpora contained a text corpus of one hundred thousand sentences of the travel conversation in various situations such as hotel, restaurant, shopping, transportation entertainment, and so on. The corpora also contained a general conversational expression corpus comprising expressions specific to oral communication. We trained a word bigram and a word trigram model using these corpora. We took advantage of the knowledge of the travel domain entities and assigned a domain semantics-based class to each word, yielding class bigram to be used in trigram smoothing. In order to supplement the insufficiency of the domain corpora, general linguistic knowledge was incorporated. The words appearing in high frequency in general text were added to the dictionary. Morphological class was defined for each word, and is used to judge if a word juncture was morphorogically allowed.

The speech recognition module was evaluated using native speakers' clean (without filler words) utterances in speaker-independent mode (gender-dependent). Test set perplexity was 26.8 for English and 33.0 for Japanese. In English evaluation, 88.9 % of the words were correctly recognized for 20 speakers where each uttered 190 sentences (6 word length on average). The word accuracy after penalizing insertion errors was 85.4%. In Japanese evaluation, 96.2% of the words were correctly recognized for 20 speakers where each uttered 200 sentences (6.8 word length on average). The word accuracy was 95.8%.

# 4. TRANSLATION

In translation of conversations, a translation module is required to cope with highly word-specific phenomena, including various colloquial and idiomatic expressions. Handling idiosyncratic word behavior is also important to improve the translation quality for the target domain. In addition, translation module is required to cover a wide range of input sentences.

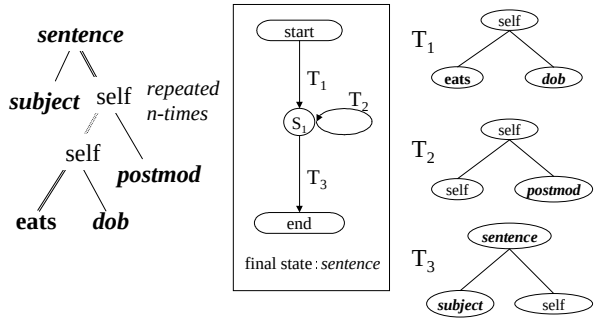
To achieve both broad coverage for general input and high quality for the target domain, we employed a rule-based method that allows writing of both general abstract rules and example-like concrete patterns in a unified framework. Precisely we adopted a strong lexicalization approach to the grammar [5,6] where all grammar rules(trees) are associated with at least one word, making all the rules lexical rules.

Furuse et al.[7] proposed an approach to spoken-language translation based on pattern matching on the surface form, combined with an example-based disambiguation method. Since our approach is more rule-oriented, we believe it is suitable to build a system with broader coverage.

## 4.1. Lexicalized Tree Automata-based Grammars

We have developed a new strongly lexicalized grammar formalism that we call Lexicalized Tree AutoMata-based Grammar (LTAMG)[8], which lexicalizes (part of) tree operations as well as the trees themselves. In this formalism, each word has a tree automaton (tree accepter) that describes how to combine the elementary trees to get the whole set of trees associated with that word. This lexicalized tree automata (LTA) allow powerful and flexible control over the tree growth. Even the complex pattern-like trees having variable part inside can be easily described by the LTA without considering side effects to other words.

Another advantage of the method is use of a simple chart-parsing algorithm which is a straightforward extension of the context-free grammar case.




(a) tree growing from "eats" (b) tree automaton (c) elementary trees

Figure 3. LTA of "eats"

Figure 3 shows an example LTA for a verb "eat". The tree growing from the verb has a form in (a). This tree is described precisely by elementary trees(c) and a tree automaton(b). The tree automaton accepts a sequence of the elementary trees. The tree(a) is presented as a sequence made by accepting $T_1$ once, then $T_2$ an arbitrary times, then $T_3$ once.

In addition to the case of example-like patterns, the formalism has enough descriptive power to develop a general grammar.

The general knowledge is represented as a shared part of the lexicalized grammar, while the domain-specific or word-specific knowledge is encoded as an individual lexicalized grammar in the dictionary.

## 4.2. Translation Module and the Grammar

Figure 4 is an overview of the translation module. Translation proceeds following the three steps: analysis, transfer and generation. The input is a sequence of words. The translation module loads the dictionary and the tree automata of the words. Some of the automata are retrieved from the rule sharing lexicon (the rule templates and the shared rules). The module carries out morphological analysis and builds a chart structure. Then the module carries out syntactic analysis to build the syntax trees. Next, this tree is transformed into a syntax tree of the target language. Finally, the words in the tree are collected, linearized, inflected and then sent to the speech synthesis module.

The translation method is based on the syntax directed translation. Each elementary analysis tree in each word is paired with an elementary generation tree. In the translation process, starting from the root, at each node of the analysis tree, the direct descendant nodes are reordered in the generation tree, according to this tree-to-tree pair. This method allows easy description for translation of example-like patterns.

We have developed the grammar rules and the dictionary to handle various phenomena that appear in conversations of the target domain, as well as rules for general phenomena. When

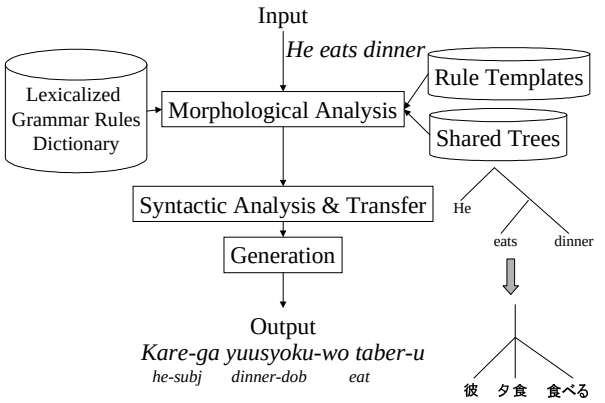



Figure 4. Translation Module

translating from Japanese to English, a missing subject, a common phenomenon in conversation, is inferred from information such as the sentence type, the polite expression types and the auxiliary verb types. When translating in the opposite direction, the speaker role and syntactic information is

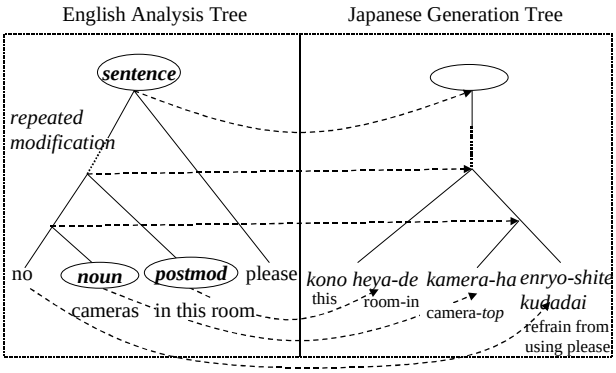



Figure 5. Translation Example

used to determine whether a polite form should be used and whether the subject should appear in the output.

Figure 5 is a translation example of a sentence containing a pattern, “No … , please”. Tree-to-tree correspondences are shown by the dotted arrows. All the tree pairs used here are contained in the LTA of the word "No", minimizing undesirable side effects to other words. Note that arbitrary number of modifiers (e.g. “in this room”) between keywords, "no" and "please", are allowed.

The translation module uses the information on the pronunciation and the pause duration, which is passed by the speech recognition module, to disambiguate word sense and to improve phrase boundary decision. The translation module also passes the phrase boundary information to the speech synthesis module to improve the synthesized speech.

Table 1 shows preliminary evaluation results of the translation for 500 travel conversation sentences. A bilingual evaluator classified the results into four categories: “Natural” (accurate translation), “Good” (there is no syntactic error and the meaning is conveyed without error), “Understandable” (loose translation but still the core meaning can be understood) and “Bad” (corrupt sentence or has an error that causes misunderstanding). The average length of the input sentences was 6.2 words for the English sentences, and 8.9 words (morphemes) for the Japanese sentences. A sentence understanding rate, that is, the ratio of the sentences other than “Bad”, of 83 – 87 % was obtained.

**Table 1: Translation Quality Evaluation**

| | Natural | Good | Understandable | Bad |
|---|---|---|---|---|
| E to J | 58.3 | 15.7 | 12.5 | 13.5 |
| J to E | 33.4 | 22.4 | 27.1 | 17.2 |

## 6. CONCLUSION

We have developed an automatic interpretation system running on a mobile PC that helps oral communication between Japanese and English speakers in various situations during their travel abroad. In order to allow a wide range of expressions and topics in the application domain, we adopted an approach which utilizes the general linguistic knowledge as well as the domain-specific linguistic knowledge. Speech recognition module performs speaker-independent large-vocabulary (50,000 Japanese words and 10,000 English words) continuous speech recognition. Translation module performs syntax directed translation based on a new lexicalized grammar rule formalism.

For the preliminary evaluation for relative short and clean sentences, a word accuracy of 85 – 96% for speech recognition and a sentence understanding rate of 83 – 87% for translation were obtained as a result. We expect that the performance will be further improved by expanding the grammar with regard to the domain-specific and colloquial expressions, which were not yet described in the grammar and caused the incorrect translation. We will also evaluate the usability of the system as aids for cross lingual communication.

## 7. REFERENCES

[1]Sugaya, F, End-to-End Evaluation in ATR-MATRIX: Speech Translation System between English and Japanese, EuroSpeech-99, pp.2431-2434 (1999)

[2]Lavie, A. et al.. JANUS III: SPEECH-TO-SPEECH TRANSLATION IN MULTIPLE LANGUAGES, Proc. ICASSP-97, pp.99-102 (1997)

[3] Reithinger, N, Robust Information Extraction in a Speech Translation System, EuroSpeech –99, pp.2427-2430 (1999).

[4]Watanabe, T et al. An experimental automatic Interpretation system: INTERTALKER, Proc. Acouts. Soc. Japan, Spring Meeting, pp.101-102 (1992) (In Japanese)

[5]Joshi, A.K et.al.. Tree-Adjoining Grammars and Lexicalized Grammars. In Tree Automata and Languages. M. Nivat and A. Podelski, ed., Elsevier Science Publishers B.V., pp.409-431. (1992)

[6]Schabes, Y. et al. Parsing Strategies with `Lexicalized' Grammars. COLING'88, pp.578-583 (1988).

[7]Furuse, O et al. Constituent Boundary Parsing for Example-Based Machine Translation. COLING-94, pp.105-111 (1994).

[8]Yamabana, K. et al. Lexicalized Tree Automata-based Grammars for Translating Conversational Texts, To appear COLING 2000 (2000)