An Experimental Multilingual Bi-directional Speech Translation System

Tomohiro Konuma[§],Kenji Matsui[†], Yumi Wakita[†], Kenji Mizutani[†], Mitsuru Endo[§], Masashi Murata[‡]
Advanced Technology Research Labs., Matsushita Electric Ind. Co., Ltd.

§ 3-10-1 Higashimita, Tamaku, Kawasaki,Kanagawa, Japan

† 3-4 Hikaridai, Seika, Souraku, Kyoto, Japan

‡ Department of Information and Communication Engineering, Osaka City University
tkonuma@mrit.mei.co.jp

Abstract

We describe an experimental Multilingual Bi-directional speech translation system utilizing small, PC-based hardware with multi-modal user interface. Two major problems for people using an automatic speech translation device are speech recognition errors and language translation errors. We focus on developing techniques to overcome these problems. The techniques include a new language translation approach based on example sentences, simplified expression rules, and a multi-modal user interface which shows possible speech recognition candidates retrieved from the example sentences. Combination of the proposed techniques can provide accurate language translation performance even if the speech recognition result contains some errors. We propose to use keyword classes by looking at the dependency between keywords to detect the mis-recognized keywords and to search the example expressions. Then, the suitable example expression is chosen using a touch panel or by pushing buttons. The language translation picks up the expression in the other language, which should always be grammatically correct. Simplified translated expressions are realized by speech-act based simplifying rules so that the system can avoid various redundant expressions. A simple comparison study showed that the proposed method outputs almost 2 to 10 times faster than a conventional translation device.

1. Introduction

Automatic speech translation is still very difficult to achieve even if the domain is limited. This is because both speech recognition errors and language translation errors cannot be avoided. On the other hand, phrase books for travelers, which have fixed expressions, are often used to assist both travel conversations and language learning. Therefore, an automatic speech translation device can be used as a replacement of the phrase books, if the device is portable and quickly retrieves the phrases. An experimental Japanese / English, Japanese / Chinese bi-directional speech translation system has been built by integrating the speech and language processing technologies developed at the authors' research laboratories. The domain is limited to frequently used travel conversations. To deal with mis-recognized input, the translation is carried out using an example-based method. Various example-based speech translation methods have been widely used and their effectiveness for spoken language processing has been confirmed [1,2,3]. However, these methods have basic problems:

(a) Even if we can limit the number of tasks (*i.e.*, the semantic domains to be covered), collecting all of the necessary expressions is difficult.
(b) Even small number of mis-recognized keywords causes totally different translation results.
(C) Users cannot recognize if the translated result is correct or not.

To solve the above problems, we propose the following approaches:

(1) We first categorize all of the keywords into classes, then the example expressions are re-written using the class symbols instead of the keywords. Also, the dependency structure of categorized keyword classes are defined and used for

measuring the confidence of each recognized keyword. Then, the possible example expressions are retrieved using the reliable keyword classes.

(2) A bilingual corpus is rewritten to yield a set of simplified expression patterns by omitting the redundant keywords in order to reduce translation errors.

(3) By asking user to select the closest example expression, he or she can confirm the meaning of the output.

Figure 1. Overview of the speech translation platform

In section 2-4, we give an overview of our speech translation system based on example expressions. In section 5-7, we describe the translation process, the keyword clustering definition of the dependency structures, and the simplified translation method. In section 8, we describe the user interface design. Finally, we report evaluation results in section 9.

2. System Overview

The key software components of the J/E and J/C speech translation system are: English, Chinese, and Japanese continuous speech recognizers, text-to-speech synthesizers, and J/E, J/C language translators and a multi-modal user interface. The system was implemented on a PC based experimental platform, equipped with a 60mm×80mm LCD display, touch panel, and USB audio input / output devices. The size is 180mm×120mm×45mm. Figure 1 gives an overview of the translation platform. There are about 1000 example expressions, covering simple travel conversations, about hotels, transportation, restaurants, shopping, and other topics of interest to travelers. The system uses 4000-word dictionary.

3. Automatic Speech Recognition

The English, Chinese and Japanese automatic speech recognizers (ASR) have been developed independently from the speech translation tasks. Those systems are speaker-independent,

medium vocabulary, continuous speech recognition system with perceptually real-time operation. All of the recognizers were designed to be small memory size applications.

In the case of Japanese speech recognizer, for example, the acoustic model is based on time-spectral vectors as feature parameters, and a statistical distance measure. The language model is based on word bi-grams trained by the task domain database. Memory reduction was achieved by reducing the number of feature parameters and by employing one-pass beam-search with two-stage pruning in the decoder. Figure 2 is shown our two-stage decoder outline. One stage in intra-word, Candidates under the score threshold are pruned frame by frame. On Figure2 (1) "AI" and "I" candidates are pruned, because their score are not satisfied the score threshold condition. "A Ku Se Sa Ri: Accessory" and "A Ku Se Su: Access" candidates are not pruned. Another stage, on end of a word general decoder is concatenated the huge number of next words, the number of next word is related vocabulary size. Our decoder for small memory size application reduce amount of calculation by restricting the number of candidates on end of a word. Candidates under the rank threshold are pruned. On Figure 2 (2) "Allergy" and "Aspirin" are pruned, because their rank are not satisfied the rank threshold condition. "Accessory" and "Access" candidates are not pruned. Out two-stage pruning decoder reduces amount of calculation by 1/4 keeping high word accuracy than only score pruning one.

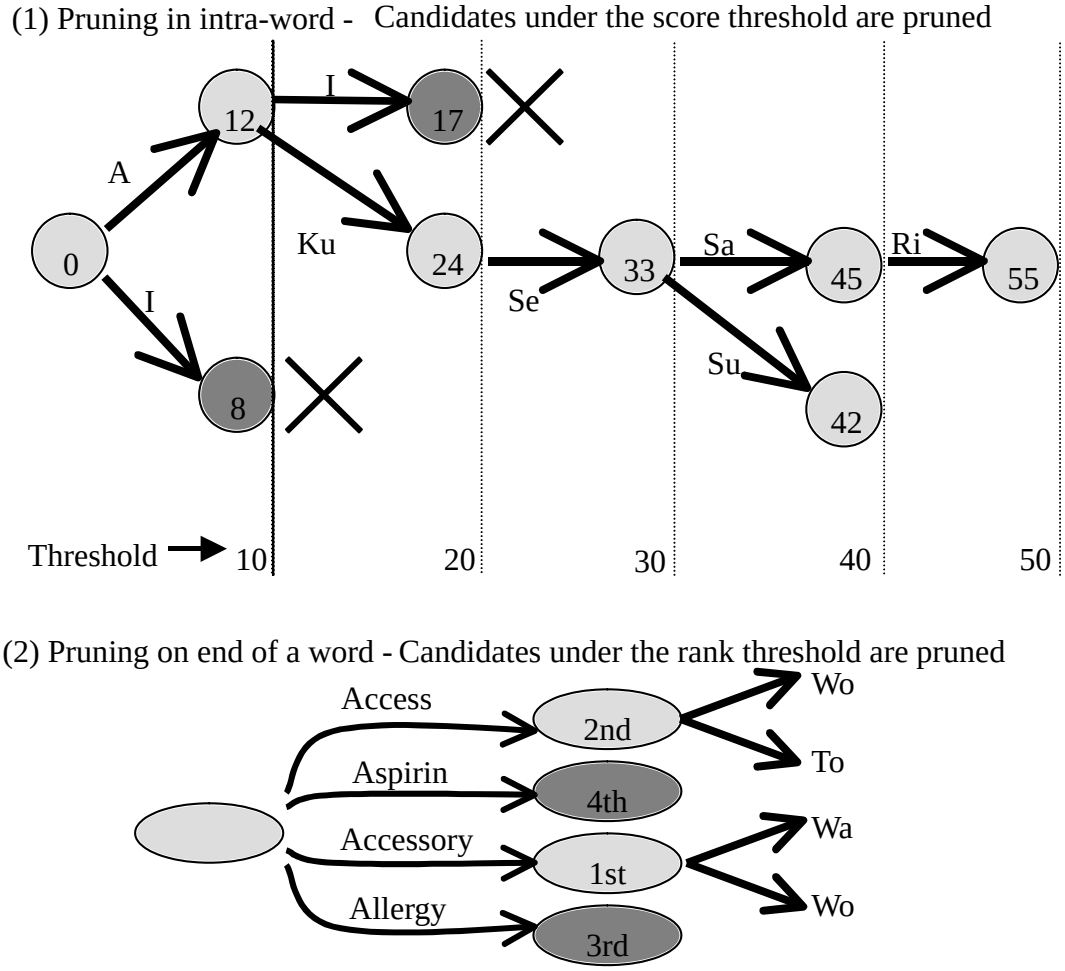

Figure 2. Our two-stage pruning decoder outline

4. Speech Synthesis

The English, Chinese and Japanese text-to-speech modules (TTS) have been also developed independently from the speech translation task. English, Chinese and Japanese TTS utilize di-phone, CV and VCV waveform concatenation units, respectively. All of the TTS modules were also optimized for small applications by compressing the concatenation units.

5. Translation

Figure 3 shows the structure of our speech translation system.

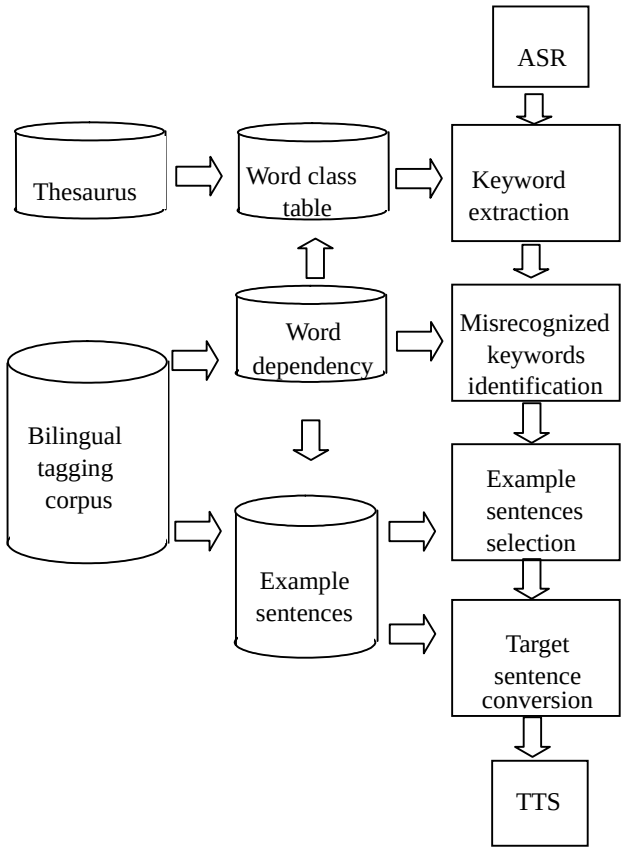

Figure 3. Block diagram of the speech translation method

The language translation is carried out using an example-based approach. The ASR output is analyzed and possible keywords are extracted. Then, the confidence for each keyword is measured by considering the dependency between keywords. The keywords that have low confidence are identified as mis-recognized words among the keywords in recognition results. The example expressions are selected from a fixed number of pre-stored sentences by looking at the dependency between keywords in the recognition results. The system picks up several candidates from the example expressions. When user chose the closest source expression, the system converts it into phrases in the target language and the TTS module generates speech output. The strength of the dependency between recognized keywords is evaluated by looking at the dependency structure extracted from the example expressions. If a recognized keyword

is not correlated with the others, the keyword is identified as an error. Thus a revised set of keywords is used for selecting example expressions.
The second key feature is a set of simplified expression rules. The goal of automatic speech translation is not to provide a perfect translation such that might be produced by a well-trained human interpreter, but to translate well enough that the correct meaning is communicated from one person to another. Our translation method maps simplified expressions from one language to another instead of doing deep parsing. In Figure 3, Target sentence conversion block use these mapped pairs. All training sentences in the bilingual corpus are grouped by rules and rewritten as simplified expressions. In addition to these strengths, the simplified expression patterns can reduce the memory size and processing time, which is a requirement for developing portable systems [4].

6. Keyword Clustering

Typical phrase book for travelers has only about 1000 fixed expressions. However, the ratio of vocabulary size related to the simple expressions is very large. We categorize the keywords into several suitable keyword classes so that the example sentence search can be done efficiently by looking at the classes instead of the actual keywords. For example, the sentence "I would like to have a coffee", could be "I would like to have [drink]." A thesaurus can be used to define keyword classes, but there are the following problems when we use the usual thesaurus classes.

(1) One keyword may have several semantic classes.
(2) Sometime a word has a different meaning from the words in the same class.

Thus, we propose a task-dependent clustering method [5]. Figure 4 shows our off-line keyword clustering method diagram. After the clustering, we use these keyword classes on translation.

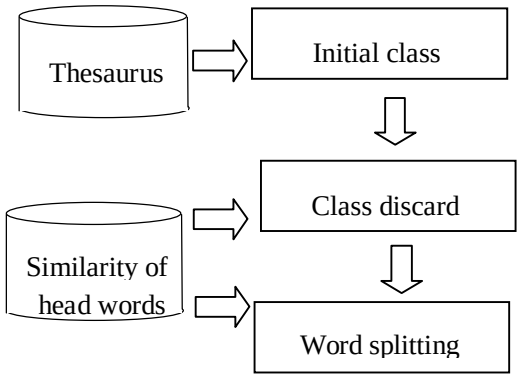

Figure 4. Task dependent keyword clustering method

First, the initial keyword classes are defined by using the thesaurus information. Next, the initial classes are re-defined according to the following conditions.

（Condition 1）The average of the similarity value between heads is calculated for each class. The classes with averages below a threshold are discarded.(Figure 4. Class discard block)

（Condition 2） If all of the similarity values between a head word and other heads are below the threshold value, the head word is split from the class. (Figure 4. Word splitting block)

This similarity is defined by the dependency analysis results of example sentences. The dependency is an asymmetric binary relationship between a word called head and another word called modifier [5]. The clustering is done to only the head side of words. When two different heads frequently depend on the same modifiers, these heads are clustered into one class. As a result of clustering, (1) suitable classes for a limited domain can be selected from

all thesaurus classes, and (2) the words whose meaning are different from the meanings of the other words can be split from the original thesaurus class. This similarity is calculated by using the following formula. If both heads, Wj and Wi, belong to class A, the similarity between Wj and Wi is calculated as follows. When two different modifiers depend on the same heads frequently, the similarity between these modifiers are regarded as high.

$$\mathrm{Sim}(W_j^A \bullet W_i^A) = \sum_{k=1}^{K} R(W_j^A \bullet Z_k) \times \sum_{k=1}^{K} R(W_i^A \bullet Z_k)$$

$$R(W_j^A \bullet Z_k) = \frac{FreqPair(W_j^A, Z_k)}{Freq(W_j^A)}$$

Where

if $R(Wi^A \bullet Zk) = 0$, then $R(Wj^A \bullet Zk) = 0$

if $R(Wj^A \bullet Zk) = 0$, then $R(Wi^A \bullet Zk) = 0$

$\mathrm{Sim}(Wj^A \bullet Wi^A)$: similarity value between Wj and Wi
Zk : k-th words of modifiers
$\mathrm{FreqPair}(Wj^A, Zk)$: Number of case Wj^A that depends on Zk
Freq(Wj) : Frequency of Wj
K : Number of all modifiers

7. Simplified Expression

In our example selection method, it is effective to use a bilingual corpus of simplified expressions to reduce the example selection errors. For example, the meaning of a typical sentence " I dropped my fork, please get me another one." has almost the same meaning as the simple sentence "Please get me a fork." From the typical sentence, these dependencies between keywords are trained: "fork" depends on "drop", and "another one" depends on "get". When this typical sentence is spoken and the keyword "fork" is mis-recognized, only the dependency between "get" and "another one" can be understood and only the following parts "Please get me another one" are selected. The meaning of the selected part is different from that of the spoken sentence.

We prepared simplified expression before using on translation. We rewrote most sentences in the bilingual corpus to simplified expressions. To reduce this selection error, we rewrote each sentence in the bilingual corpus to a simplified expression by omitting redundant expressions and changing to a compact pattern that can be correctly understood, such as "Please get me a fork". Furthermore, the simplified expressions seem to be effective in reducing the size of translation rules and target sentence generation rules. This is essential for developing a portable translation system.

To rewrite typical expressions into simplified expressions, we have following principles:

(a) The same expression patterns are used for same speech-act to reduce resource requirements.

(b) Omitting the redundant expressions reduces the translation errors

In travel conversation, many sentences whose speech-acts are "request", or "confirmation", or "submission" can be found. For these speech-acts, the expression patterns in Table 1 are used.

Speech-act	Simplified principles	Example
Request	NP + please VP + please	Coffee, please
Confirmation	Any + NP	Any painkiller?
Submission	NP + okay	Reservation, okay ?

Table 1. Examples of simplified expressions

In addition to the expressions of Table 1, we decided on the expressions for speech-acts, such as “question” and “negation”. The number of main expression patterns became eleven in total. Using these expression patterns, we rewrote the bilingual corpus of the travel conversation domain. As a result of rewriting, 78% of the sentences matched the rules for simplifying: The 60％ of the sentences matched principle (a), and 18% of the sentences could matched principle (b). The number of example sentences can be decreased to 72% of the original number by using simplified expressions.

8. User Interface

Figure 4 shows an example of the speech translation result, with: (1) the speech recognition result, (2) possible example sentences retrieved by the keywords, and (3) the translated text. Also, an agent is showing the status of the translation process, such as idling, translating, and finish translating. The user, first, selects the translation mode (the choices are J-to-E, E-to-J, J-to-C, or C-to-J). When the system accepts the mode command, the system generates a beep sound. The user must speak to it right after the beep. Then, the user can see the recognition result in the bottom window, and possible example expressions in the upper window. By touching the most suitable expression, the system translates it to the target

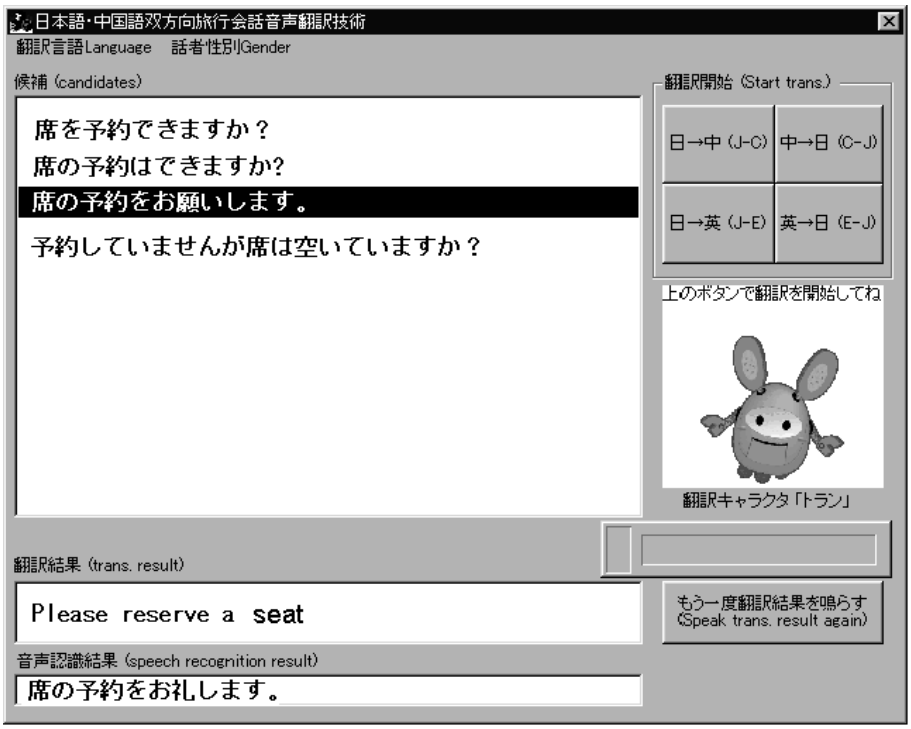

Figure 5. An example of the speech translation result

The upper windows: (The 1st line and the 2nd line) “Can I reserve a seat?” ; (The 3rd line) “Please reserve a seat.” ; (The 4th line) “I have no reservation, but is there a seat available?”. The bottom windows : “Thank you for making a seat reservation.”, instead of “Please reserve a seat.”, because of an recognition error.

language shown in the middle window. The synthesized speech comes out when the user touches the “SPEAK” button. Since the example sentences come from a bilingual corpus, the user can trust the translated outputs. In Figure 4, there is an error in the ASR result. However, even if the system makes the error, the upper window shows a possible candidate because the system is using only the reliable keywords for retrieving the example sentences.

9. Evaluation

The goal of automatic speech translation is not to provide a perfect translation by well-trained human interpreter, but to translate quick and enough that the correct meaning is communicated from one person to another.
A comparison test was conducted to assess if the proposed speech translation method retrieved faster than a conventional portable translation device. The conventional device, we evaluated, has a straightforward hierarchical search method, which has 8 categories. Three subjects participated in this study. They tried 6 sentences to obtain the translated output, and measured the total processing time for both systems. Table 2 shows the average retrieval time in the case of conventional method. The average retrieval time for the speech translation system described in this paper was 5 seconds.

Text	Processing time
Give me a receipt, please.	58
Non-smoking sheet please.	75
Express mail please.	12
Please give me a discount.	39
I lost my wallet.	25
How much is it?	22

Table 2. Average retrieving time by 3 subjects in the case of conventional method (sec.)

10. Conclusion

A simple comparison study showed that the proposed method outputs almost two to ten times faster than the conventional translation method. Also, a combination of the proposed techniques can provide accurate language translation performance even if the speech recognition result contains some errors.

11. Acknowledgement

Our thanks to Chinese Academy of Science Institute of Automation for allowing us to use their ASR and NLP technologies. We also thank Panasonic Speech Technology Laboratory for their participation in the project.

12. References

[1] O.Furuse,H.Iida: Constituent Boundary Parsing for Example-based Machine Trasnlation. *Proc. Coling94,* pp. 105-111 (1994)

[2] Satoshi Sato, MBT2: a Method for Combining Fragments of Examples in Example-based Translation. *Artificial Intelligence* 75, pp.31-49 (1995).

[3] K.Ishikawa, E.Sumita, and H.Iida: Example-Based Error Recovery Method for Speech Translation. *ICSLP98*, pp.1147-1150 (1998)

[4] Chengqing Zong, Yumi Wakita, Bo Xu, Kenji Matsui and Zhenbiao Chen : Japanese-to-Chinese Spoken Language Translation Based on the Simplified Expression. *ICSLP2000*

[5] Yumi Wakita, Kenji Matsui, and Yoshinori Sagisaka: Fine Keyword Clustering using a Thesaurus and Example Sentences for speech translation. *ICSLP2000*