Response Generation based on Statistical Machine Translation for Speech-Oriented Guidance System

Kazuma Nishimura, Hiromichi Kawanami, Hiroshi Saruwatari and Kiyohiro Shikano*
* Graduate School of Information Science, Nara Institute of science and Technology, Japan
E-mail: {kazuma-n,kawanami,sawatari,shikano}@is.naist.jp

***Abstract*—An example-based response generation is a robust and practical approach for a real-environment information guidance system. However, this framework cannot reflect differences in nuance, because the set of answer sentences are fixed beforehand. To overcome this issue, we have proposed response generation using a statistical machine translation technique. In this paper, we make use of N-best speech recognition candidates instead of manual transcription used in our previous study. As a result, the generation rate of appropriate response sentences was improved by using multiple recognition hypothesis.**

I. Introduction

Automatic speech recognition (ASR) has been widely applied to dictation, Voice Search, and car navigation, to name a few. In this paper, we describe a speech-oriented information guidance system, *Takemaru-kun*[1], aimed at realizing a natural speech interface using ASR.

Takemaru-kun is a real-environment speech-oriented information guidance system whose task domain is not given before the operation starts. As *Takemaru-kun* employs an example-based question answering system, responses to users' questions have been added on demand.

A response to a users' question is selected by referring to a question and answer database (QADB), which can be easily maintained without paying particular attention to the scope of the system.

One of the problems in an example-based system such as *Takemaru-kun* is that the response sentences cannot reflect differences in nuance as the set of answer sentences are fixed beforehand. To realize a familiar speech interface, it is preferable to arrange responses using more appropriate phrases.

We have already proposed an approach to generate response sentences by introducing a Statistical Machine Translation (SMT) technique[2]. In this paradigm, we treat the question set and the answer set as different languages. That is to say, a question input to the system is "translated" to the corresponding system answer by SMT models.

In the previous work, we used only manual transcription of users' utterances as the training data and also as the test data. In this study, we treat N-best speech recognition candidates as system input and propose a method of improving the performance of SMT-based response generation. In this work, experiments are conducted using real users' Japanese utterances.

II. SMT-based response generation

A. Statistical machine translation

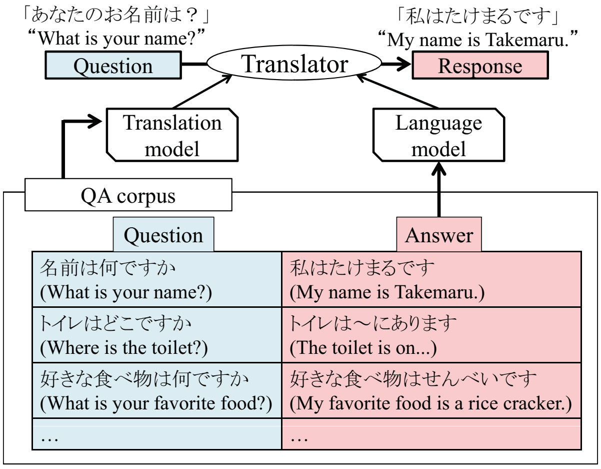
SMT builds statistical translation models from the analysis of bilingual corpora and enables the translation of a language into another language automatically. Suppose you want to translate a sentence from source language $\mathbf{f}$ into a sentence of target language $\mathbf{e}$. There are innumerable choices of translated results $\mathbf{e}$. The decoder in the SMT system calculates $P(\mathbf{e}|\mathbf{f})$, the probability that $\mathbf{e}$ is the translation result of $\mathbf{f}$, for all pairs of $(\mathbf{e}, \mathbf{f})$. The system outputs the sentence $\hat{\mathbf{e}}$ for which $P(\mathbf{e}|\mathbf{f})$ is the greatest. We can express this problem with the following log-linear model.

$$\begin{aligned}\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) \\ &= \arg\max_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}) \qquad (1)\end{aligned}$$

Now, M, λ_m, and h_m denote the total number of features, the weight of each feature, and each feature function, respectively. Examples of feature functions are the translation model and language model. The translation model gives the probability of translating, and the language model is the expression of the fluency of the sentence. The translation model is built from the analysis of bilingual corpora and the language model is built from the analysis of the corpus of the target language. The IBM model[3], built by learning the alignment of words, was used originally as the translation model. Recently, a phrase-based translation model has been proposed[4]. In the phrase-based model, a phrase is used as the alignment unit instead of a word. Herein, "phrase" means simply a sequence of words, not a linguistic unit, for example, verb phrase or noun phrase. In this study, we apply the phrase-based SMT.

B. Adapting SMT to response generation

In our approach, we assume that question sentences can be translated into response sentences if we consider questions and response sentences as different languages. Figure 1 illustrates how the technique is applied to response generation. In the language translation task, translation models are built from bilingual corpora, for example, English and French. Now, in the response generation task, translation models are built from QA pairs, which are manually maintained and has been used



Question	Answer
名前は何ですか (What is your name?)	私はたけまるです (My name is Takemaru.)
トイレはどこですか (Where is the toilet?)	トイレは〜にあります (The toilet is on...)
好きな食べ物は何ですか (What is your favorite food?)	好きな食べ物はせんべいです (My favorite food is a rice cracker.)
…	…

Fig. 1. SMT-based response generation.

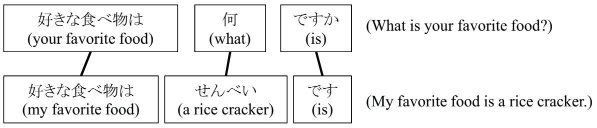


Fig. 2. Alignment expected to be learned.

in the real *Takemaru-kun* system. Language models are built from answer sentences. The following procedures are the same as the language translation task. In the SMT-based response generation, it is expected that an alignment such as that in Fig. 2 is trained in the training process.

III. SMT response generation using ASR candidates

In our previous work, we evaluated the performance of response generation using SMT with manual transcription of users' utterances. In actual operation, speech recognition candidates are used as inputs to the response generation module. Multiple candidates are used to select the most similar example question in the QADB.

In general, speech recognition candidates include recognition errors, which may involve a decline of response performance when using a translation model from the transcriptions. Therefore, we introduce ASR candidates also in building the translation model. We expect to be able to obtain the translation model considering recognition errors.

A. A method using multiple recognition candidates

In the speech recognition process, a recognition engine outputs multiple recognition candidates. We propose a method in which these candidates are made use of.

In the training phase, N-best recognition candidates are separated one by one, and their response sentence is connected to each candidate (Fig. 3). This process realizes the extension of training data as practical input.

In the generation phase, each N-best candidate is translated into a response sentence candidate. The response sentence that obtains the highest translation score is used as the final output. This process is illustrated in Fig. 4. This method enables the generation of more appropriate response sentences.

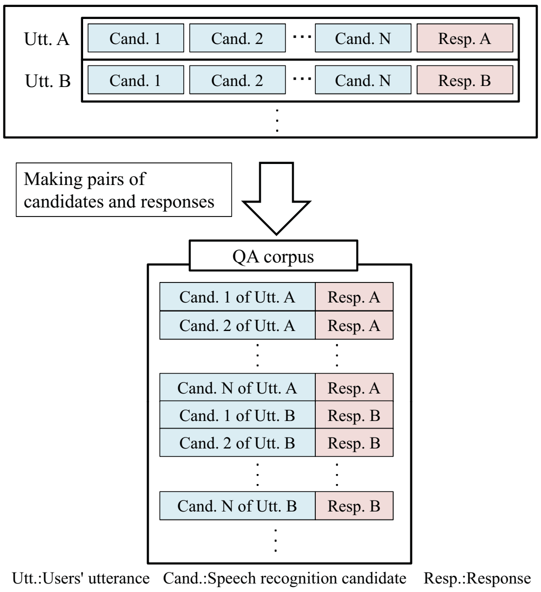


Fig. 3. Proposed method in training phase.

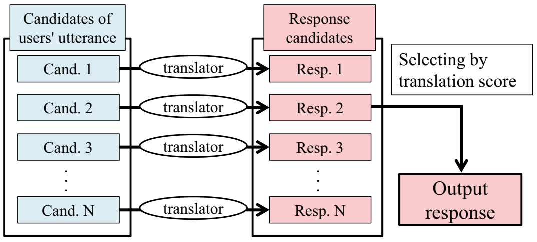


Fig. 4. Proposed method in generation phase.

IV. Experiments

A. Experimental condition

We employed a dataset that consists of speech recognition outputs of adult users' utterances and the answer sentences tagged on them. The ASR engine is Julius 4.2[1]. The number of the kinds of answer sentences in the QADB is 276. The domains of responses are information about, for example, facilities or sightseeing, chatting, and greeting. The dataset was collected with the *Takemaru-kun* system from Nov. 2002 to Oct. 2004 (Table I). As preprocessing, the dataset were tokenized by ChaSen[2].

We built the translation model from these QA pairs, and built the language model from the answer sentences, excluding the pairs of Jul. and Aug. 2003. When training with a single candidate (1-best), the training data consist of 18509 pairs.

[1] http://julius.sourceforge.jp

[2] http://sourceforge.jp/projects/chasen-legacy

TABLE I
DATASET FEATURES

Training data	Period	Nov. 2002-Oct. 2004 (excluding Jul. & Aug. 2003)
	# of data	18509 pairs(1-best) 184983 pairs(10-best) 912289 pairs(50-best)
Development data	Period	Jul. 2003
	# of data	872 pairs(1-best)
Test data	Period	Aug. 2003
	# of data	959 utterances
Word Accuracy		86.88%

Extended data with N-best candidates consist of 184983 pairs (10-best) and 912289 pairs (50-best).

The data of Jul. 2003 were used as development data, and the feature weights were optimized by minimum error training[5]. The development data consist of only 1-best candidates.

The data from Aug. 2003 were used as test data. Out-of-Task utterances are excluded. N-best recognition candidates were first translated into response candidates. Then the response which had the highest translation score was used as a final result. Experiments using 1-best, 10-best, and 50-best candidates as input were conducted. The word alignment was obtained by running GIZA++(http://code.google.com/p/giza-pp/), and the 3-gram language model built by SRILM[3], and extracted phrases and decoded sentences by Moses using the default settings.

B. Criterion

We evaluated the results subjectively by one native student from the viewpoint of "appropriateness" as a response. "Appropriateness" consists of the following two factors.

- informativeness
  (the sentence includes necessary information)
- naturalness
  (the sentence is natural in a language)

First, generated sentences were manually judged whether they were informative and natural separately. Sentences which are informative and natural are labeled as "appropriate." Experimental condition used to generate each sentence was not announced to the evaluator.

C. Results

The rate of appropriate responses is shown in Fig. 5. The horizontal axis is the number of input candidates.

In the case using translation model built by transcriptions, 59.6 % of test sentences were appropriate [2]. This value is used as a baseline. As a reference, the response accuracy of the conventional example-based method using manually transcribed QADB is 82.3 %.

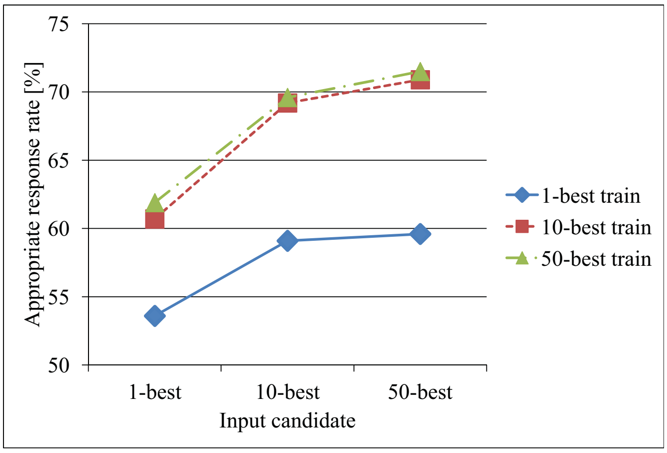


Fig. 5. Generation rate of appropriate responses.

When training data consist of 1-best speech recognition candidates and input data consist of 1-best candidates, 53.6 % of sentences were appropriate. From the viewpoint of the number of input candidates, the results of 10-best input and 50-best input were superior to that of 1-best. Considering the number of training data candidates, the results when using multiple candidates were superior, as well. In particular, the results of 50-best input and 50-best training reached 71.5 %. It is supposed that increase of the number of recognition candidates avoids to lose linguisitically appropriate candidates.

Figure 6 shows the rate of informative response and Fig. 7 shows the natural sentence rate. The more candidates are used in training data and input data, the more the natural sentence rate is improved. However, the informative sentence rate was not improved compared with naturalness. The reason of this phenemenon is assumed that increasing N-best candidates mainly contributes to variations of literal expression, which leads to improvement of naturalness.

V. DISCUSSION

The more candidates were used as training data, the more the number of sentences generated appropriately. This might be because a translation model that absorbed the recognition error or fluctuation could be built. In the example shown in Fig. 8, speech recognition of the question was incorrect. When training with 1-best candidates, it was impossible to translate "sanara", which is a recognition error of "sayo:nara" which means "Good-bye," because it did not exist in training data. However, as "sanara" existed in the training data of 10-best candidates, translation was successful.

The more candidates are used as input data, the more the number of sentences generated appropriately. It is hypothesized that the responses that do not have excrescent words were selected owing to their higher likelihood of the language model. The example shown in Fig. 9 is a translated response sentence generated from the 4th candidate, the translation score of which is the highest among the 10-best candidates. A filler, "etto" is translated to "washitsu" which means "Japanese room" in the example of the 1st candidate. On the other hand, the 4th candidate does not include fillers and the translated

[3] http://www.speech.sri.com/projects/srilm/

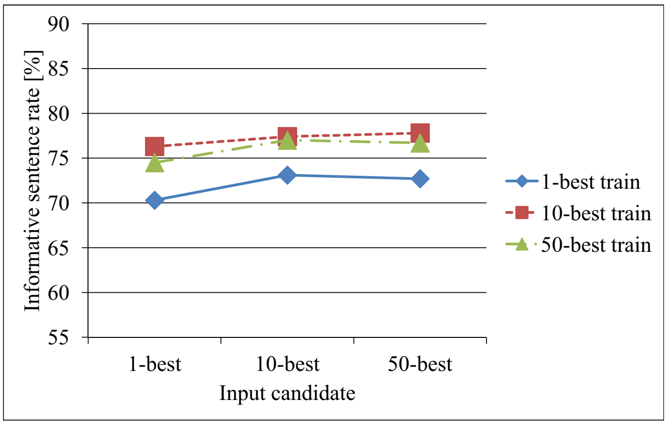


Fig. 6. Generation rate of informative sentences.

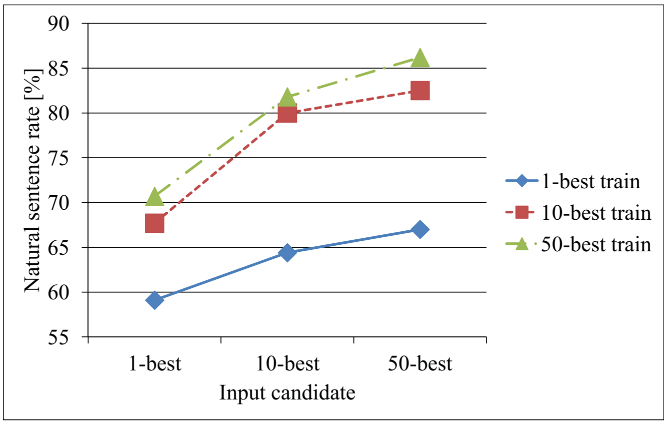


Fig. 7. Generation rate of natural sentences.

sentence realizes an appropriate response with the highest translation score.

It is assumed that the rate of natural responses increases owing to the effect of these factors.

VI. Conclusions

In this paper, we proposed an SMT-based method of response generation using multiple recognition candidates and conducted an experiment using ASR candidates. The proposed method was effective in generating appropriate responses. This method improved the rate of natural sentences, and contributed to the improvement of appropriate response rate. However, the informative sentence rate did not improve very much. In future work, the informative sentence rate must be improved.

VII. Acknowledgements

This work was partially supported by CREST (Core Research for Evolutional Science and Technology), Japan Science and Technology Agency (JST).

INPUT:Sanara
(A recognition error of “Sayo:nara (Good-bye)”)
OUTPUT:
1-best training:Sanara
(The input phrase which could not be translated)
10-best training:Sayo:nara mata yoroshiku onegaishimasu
(In English “Good-bye, see you again.”)

Fig. 8. Example of effect of training with multiple candidates.

Generation from the 1st candidate:
INPUT:Etto teNkiyoho: oshiete kudasai
(“Well, please tell me weather forecast.”)
OUTPUT:Washitsu teNkiyoho: no ho:mupe:ji ni akusesu shimasu
(“Japanese room I’ll show you a web site of weather forecast.”)

Generation from the 4th candidate (selected):
INPUT:TeNkiyoho: oshietekudasai
(“Please tell me weather forecast.”)
OUTPUT:TeNkiyoho: no ho:mupe:ji ni akusesu shimasu
(“I’ll show you a web site of weather forecast.”)

Fig. 9. Example of the response generated from the 4th candidate being selected.

References

[1] Ryuichi Nisimura *et al.*, “Public speech-oriented guidance system with adult and child discrimination capability,” *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1, pp.433–436, 2004.

[2] Kazuma Nishimura *et al.*, “Investigation of statistical machine translation applied to answer generation for a speech-oriented guidance system,” *Proc. of APSIPA Annual Summit and Conference,* 2011.

[3] Peter F. Brown *et al.*, “The mathematics of statistical machine translation: parameter estimation,” *Computational Linguistics,* 19(2), pp.263–311, 1993.

[4] Philipp Koehn *et al.*, “Statistical phrase-based translation,” *Proc.of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* 1, pp.48–54, 2003.

[5] Franz Josef Och, “Minimum error rate training in statistical machine translation,” *Proc. of the 41st Annual Meeting on Association for Computational Linguistics*, Vol. 1 of *ACL ’03*, pp. 160–167, 2003.