Combination of 3 Types of Speech Recognizers for Anaphora Resolution*

Kazutaka Shimada, Noriko Tanamachi, and Tsutomu Endo

Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Iizuka Fukuoka Japan 820-8502
{shimada, n_tanamachi, endo}@pluto.ai.kyutech.ac.jp

Abstract. In this paper, we propose a method for anaphora resolution in speech understanding for a livelihood support robot. For robust speech recognition, we combine two types of speech recognizers; a large vocabulary continuous speech recognizer (LVCSR) and domain-specific speech recognizers (DSSR). One problem in the anaphora resolution is lack of the antecedent in the outputs. To solve the problem, we introduce 2 types of DSSRs; one medium-scale DSSR and several small DSSRs. In this paper, we describe the basic idea of our multiple speech recognizer first. The selection process in the recognizer is based on the similarity between the LVCSR and each DSSR. Then, by using the outputs from the LVCSR and the medium-scale DSSR, we resolve anaphoric expressions in the current output from a small-scale DSSR. The experimental result shows the effectiveness of our method.

Keywords: Anaphora resolution, Multiple speech recognizer, Combination

1 Introduction

Speech understanding and dialogue systems have been developed for practical use recently. These systems often recognize user utterances incorrectly. It is important to deal with speech recognition errors for speech understanding systems. Extracting keywords and understanding an utterance using them reduce speech recognition errors (Bouwman *et al.*, 1999; Komatani and Kawahara, 2000). Combining some recognizers is one of the best approaches to improve the accuracy of speech understanding systems (Isobe *et al.*, 2007; Utsuro *et al.*, 2004). Utsuro et. al. (2004) have obtained high accuracy by using some speech recognizers' outputs. However they dealt with word error reduction only. Although Isobe et. al. (2007) have proposed a multi-domain speech recognition system based on some domain-specific recognizers, their system cannot treat out-of-domain utterances such as a chat between users. However chat utterances often include significant information as the context of the dialogue.

In this paper we propose a simple and effective speech understanding method based on a large vocabulary continuous speech recognizer (LVCSR) and some domain-specific speech recognizers (DSSR). We call it "One Generalist and Some Specialists (OGSS) model". Figure 1 (a) shows the outline of the model. In our system, the LVCSR is the generalist, namely domain-independent, and the DSSRs are specialists, namely domain-dependent. We focus on the difference between outputs generated from the generalist and specialists. By using this method, we can recognize domain-dependent speech inputs with high accuracy and also handle context information in domain-independent speech inputs.

The task of this system is speech understanding for a livelihood support robot. The DSSRs recognize particular utterances about orders; e.g., order utterances from elders who need care and order utterances from nurses. We construct the grammar-based DSSRs for order utterances with

* This research was supported by NEDO, Intelligent RT Software Project, 2010.

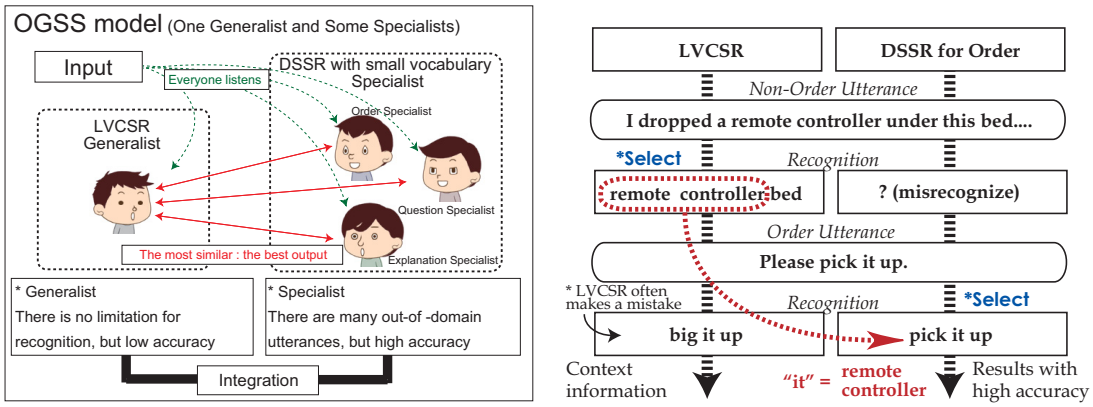

Figure 1: The OGSS model and the effectiveness

a small vocabulary and high accuracy for each order type. We use the LVCSR for recognition of utterances that the DSSRs can not recognize, such as a chat between users. The information recognized by the LVCSR is of assistance for context construction of a dialogue. If we handle these different speech recognizers selectively and integratively, we realize a flexible and robust speech understanding method. Figure 1 (b) shows the effectiveness of the proposed multiple recognizer. The DSSR achieves the order recognition with high accuracy and the LVCSR supplies lack of information in the order utterances.

In general, there are many anaphoric expressions in a dialogue. Anaphora resolution is one of the most important tasks for understanding the dialogue. In this paper, we also propose an anaphora resolution method in the multiple recognizer. By using previous outputs from the LVCSR and some DSSRs, we resolve an anaphora in the current output. For example, with respect to the utterance "Please pick it up" in Figure 1 (b), the system identifies that the word "it" in the utterance is the phrase "remote controller" which was recognized by the LVCSR in the previous utterance. The antecedent often appears in non-order utterances, that is outside of DSSRs. Therefore the target word is usually recognized by a LVCSR. However, the accuracy of the LVCSR is generally insufficient. The low accuracy of the detection of the antecedent in the speech recognition process leads to the decrease of the accuracy of the anaphora resolution process because the antecedent does not exist in the output of the speech recognizer. Here we apply a medium-scale DSSR to the multiple recognizer. It contains words of the target situation. In other words, the vocabulary of the medium-scale DSSR consists of the union of each small-scale DSSR, such as a nurse's order DSSR and a patient's order DSSR. By using the medium-scale DSSR, the accuracy of non-order utterances often improves. It leads to the improvement of the accuracy of the anaphora resolution method.

In Section 2, we explain the basic idea of the multiple speech recognizer. In other words, it is to select an output from each recognizer. In Section 3, we describe an anaphora resolution method based on the combination of 3 types of speech recognizers. Then, we evaluate the method in terms of the output selection and anaphora resolution in Section 4. Finally we conclude this paper in Section 5.

2 Combination model

2.1 Basic idea

In this section, we explain the process of output selection in the OGSS model. In this process, we focus on a difference of outputs generated from each recognizer. Even human beings tend to misunderstand words which consist of similar pronunciations (Komatani *et al.*, 2005). Here

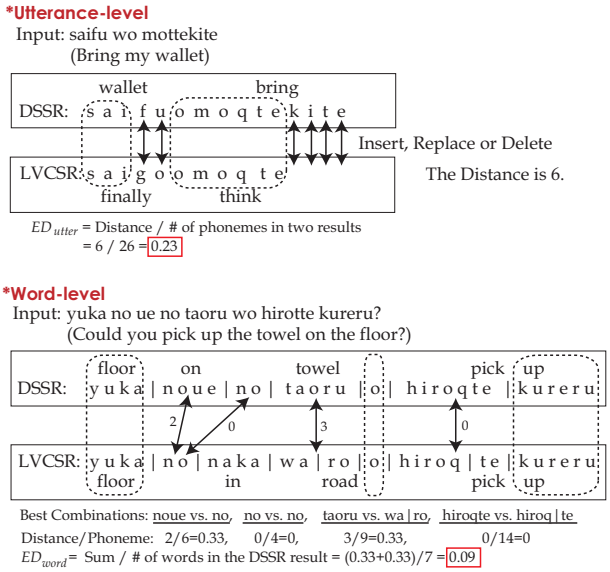

Figure 2: The edit distance calculation

we focus on the output of the LVCSR. If an input is an order utterance, a DSSR and the LVCSR generate similar outputs on phoneme-level because the LVCSR is domain independent. On the other hand, if the input is not an order utterance, they often generate different outputs even on the phoneme-level because the DSSR never generates the correct result for non-order utterances.

In this paper, we apply an unsupervised approach to the output selection method. We use the edit distance as the similarity measure. The correspondence such as the edit distance is one of the most effective measures to identify high confidence words in outputs (Utsuro *et al.*, 2004) and to extract similar word pairs (Komatani *et al.*, 2005). In our method, if an input is an order utterance, the edit distance between the outputs from a DSSR and the LVCSR becomes small. However if the input is not an order utterance, that between the outputs from each DSSR and the LVCSR becomes large. In our method, we compute the edit distance of utterance-level and word-level by using a DP matching algorithm. In the process, we compute the edit distance between phonemes of words for both levels.

The rules to judge an utterance are applied in the following order:

1. Compute the edit distance of the utterance-level (ED_{utter}) between the LVCSR and each DSSR. For the outputs of which the edit distance is less than $thresh_{utter}$, we select the output of the DSSR which contains the minimum ED_{utter} as the final output.
2. Compute the edit distance of the word-level (ED_{word}) between the LVCSR and each DSSR. For the output of which the edit distance is less than $thresh_{word}$, we select the output of the DSSR which contains the minimum ED_{word} as the final output. Otherwise, the LVCSR as the final output.

The ED_{utter} is the edit distance value on the utterance-level. The ED_{word} is the average of the edit distance value computed on word-level. These values are normalized by the number of phonemes in the outputs. The $thresh_{utter}$ and $thresh_{word}$ are threshold values for the judgment. These values are decided experimentally.

In the computation of the word-level, we eliminate word pairs that are matched completely first. Next, we compute all the combinations of the other. Finally, we employ the minimum combinations as the word-level edit distance. Figure 2 shows an example of the calculation of the ED_{utter} and ED_{word}. In the figure, the dotted line denotes completely matched words. The numerals with arrows denote the original edit distance of the word pair. In the alignment process

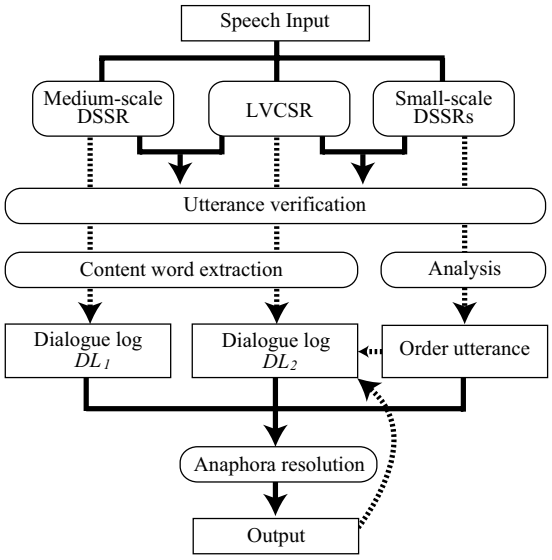

Figure 3: The anaphora resolution with the multiple recognizer.

of word pairs, we select pairs which have the minimum value of the edit distance. In other words, we admit overlap of word pairs. For example, " noue vs. no " and " no vs. no " in Figure 2.

2.2 Recognizers

The OGSS model consists of a LVCSR and some DSSR. The LVCSR is used for utterance verification, namely output selection, and capturing the context information in a dialogue. However, the accuracy of the LVCSR is generally insufficient. The accuracy is important for an anaphora resolution process. The low accuracy of the LVCSR leads to the decrease of the accuracy of the anaphora resolution process because the antecedent does not exist in the output of the speech recognizer.

In this paper, we used 2 types of DSSRs; some small-scale DSSRs and a medium-scale DSSR. The small-scale DSSRs are used for each particular domain or task; e.g., order utterances from elders who need care and order utterances from nurses. On the other hand, the medium-scale DSSR is used for capturing the context in the target situation (a livelihood support robot in this paper). In other words, it is a integrated DSSR of small-scale DSSRs. The vocabulary of the medium-scale DSSR is the union of each small-scale DSSR.

As a result, the multiple speech recognizer consists of one LVCSR, one medium-scale DSSR and some small-scale DSSRs. In the utterance verification process, our method compares the LVCSR with some small-scale DSSRs for the output selection. Also it compares the LVCSR with the medium-scale DSSR for generating a context word list with high accuracy from the medium-scale DSSR.

3 Understanding and Anaphora Resolution

In this section we explain an anaphora resolution process in the OGSS model. Figure 3 shows the outline of the anaphora resolution process. In the figure, DL_1 is a content word list detected from the output of the medium-scale DSSR. In the utterance verification process, if the output from the medium-scale DSSR is similar to that from the LVCSR, content words in the DSSR's output are stored. In the same way, if an input is out-of-vocabulary in all DSSRs, that is the large edit distance between the LVCSR and all the small-scale DSSR, the output of the LVCSR is stored to DL_2. Otherwise, the output from a small-scale DSSR is stored to DL_2.

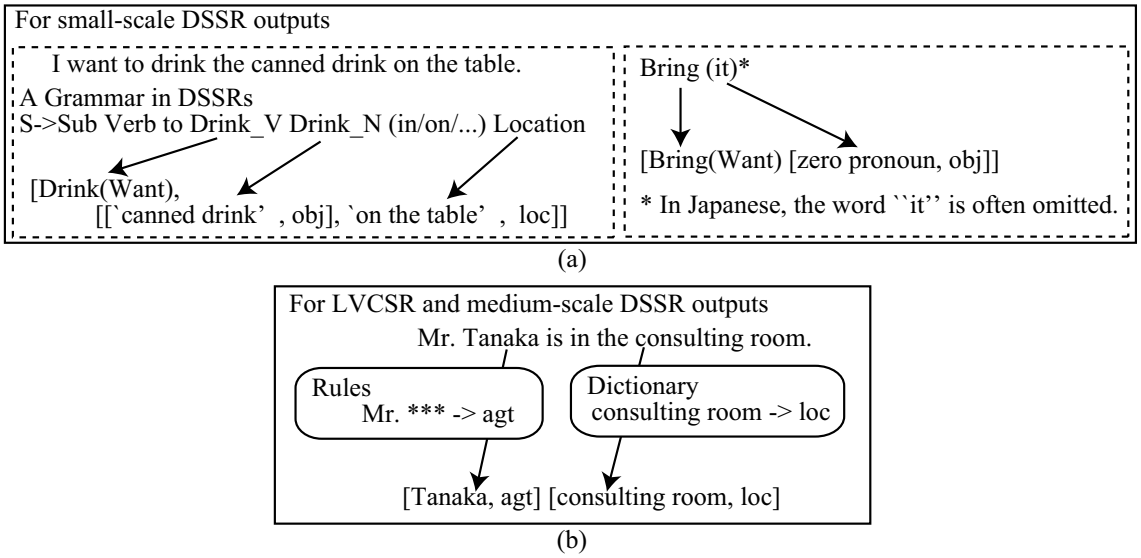

Figure 4: Examples of the output analysis

3.1 Understanding of Outputs from OGSS model

The output in the previous section, namely the output selection process, is an output of a speech recognizer. For the anaphora resolution process, we need to analyze the output.

For outputs from small-scale DSSRs, we convert them into a semantic frame. We utilize grammar information of DSSRs for the process. Each DSSR consists of 100-200 words and approximately 100 grammar patterns including approximately 50 categories. Figure 4 (a) shows an example of the grammar pattern and categories. The categories often contain semantic constraints such as "Drink_N" and "Location". In this process, we also use a dictionary which is described for required slots of each verb. We detect zero pronouns in utterances from the small-scale DSSRs by using the dictionary.

For outputs from a LVCSR, we extract keywords by using some rules based on surface expression. For outputs of a medium-scale DSSR, we also extract keywords by using the categories in the vocabulary. Figure 4 (b) shows examples of the process. In the figure, "obj", "loc" and "agt" denote case markers.

3.2 Anaphora Resolution

If an utterance contains an anaphoric expression, our system detects the antecedent from previous utterances. The anaphora resolution process is based on a scoring method for words in DL_1 and DL_2. In the scoring method, we also focus on (1) the distance from the current utterance and (2) change of situation.

First, we explain the 1st step of the scoring method; weighting of each word. For a word k_i in DL_1 and DL_2, we set the weights in the following manner. The i denotes the location of the k in the DL_1 and DL_2.

For the dialogue logs from the medium-scale DSSR (DL_1),

$$Conf^1_{k_i} = CN^{DL_1}_{k_i} \tag{1}$$

where CN_{k_i} denotes the confidence measure computed from the LVCSR or the medium-scale DSSR for each word and the range is [0, 1].

The DL_2 contains 3 types of outputs; outputs from the LVCSR, original outputs from the small-scale DSSRs and outputs from anaphora resolution. For the LVCSR and original DSSR's

outputs, we use the CN_{k_i} as the weight.

$$Conf^2_{k_i} = CN^{DL_2}_{k_i} \tag{2}$$

If the k_i is the output of the anaphora resolution process, we set a constant number.

$$Conf^2_{k_i} = 0.7 \tag{3}$$

The reason why the value is constant and small as compared with that of original outputs is that the accuracy of the anaphora resolution process is not always high, that is insufficient confidence.

Next, we compute a score for each k_i. Here we apply a decay factor based on the distance and the situation to the scoring process.

$$DF_{k_i} = \frac{1}{dist^2_k} \times sc^n \tag{4}$$

where $dist_k$ is the distance between the current utterance that contains the anaphoric expression and the previous utterance that contains the antecedent. sc is a parameter for the change of situation. We define "change of a speaker" and "change of the location of a robot" in a dialogue as the "change of situation". The "change of situation" denotes the change of the topic in conversation. n is the number of changes. If there is no change of situation for a target word k_i, the n is 0. In this paper, we set $sc = 0.1$. We multiply the $Conf^1_{k_i}$ and the $Conf^1_{k_i}$ by the decay factor DF_{k_i}.

$$dConf^1_{k_i} = Conf^1_{k_i} \times DF_{k_i} \tag{5}$$

$$dConf^2_{k_i} = Conf^2_{k_i} \times DF_{k_i} \tag{6}$$

The final score of k_i is computed as follows:

$$Score_{k_i} = \alpha \times dConf^1_{k_i} + \beta \times dConf^2_{k_i} \tag{7}$$

where α and β are weight parameters for each $dConf$. We compute scores of all candidates k_i that appear in previous N-utterances, and select the word that contains the maximum score in them. In this paper, we set $N = 10$, $\alpha = 0.5$ and $\beta = 1.5$. Here α is the weight for the outputs from the medium-scale DSSR and β is the weight for the outputs from the LVCSR and small DSSRs. In this scoring, we set a small value to α as compared with β. The reason is that the outputs from the medium-scale DSSR often contain insertion errors because there are many out-of-vocabulary words in a chat.

4 Experiment and discussion

4.1 Speech recognizer in the experiment

We used Julius as the LVCSR and Julian as the DSSR (Lee *et al.*, 2001). Julius is a famous large vocabulary continuous speech recognition decoder based on word N-gram and context-dependent HMM. In this experiment, we used original acoustic and language models. The Julian consists of a vocabulary and a grammar file. For the grammar file we describe sentence structures in a BNF style, using word category names as terminal symbols. The vocabulary file defines words with their pronunciations (i.e., phoneme sequences) for each category. Here we design grammar and vocabulary files of the Julian which accepts only specific utterances from users. In this experiment, we used 4 small-scale DSSRs that we constructed by hand. The DSSRs are as follows:

- Order Utterances from patients: e.g., "Please bring the remote controller on the table"
- Order Utterances from nurses: e.g., "Carry these meals to patient's rooms"

- System Commands: e.g., “Move to the right by 50cm”
- Question Utterances: e.g., “Where is my cellphone?”

Each DSSR consists of approximately 200 words and 100 grammar patterns.

For the medium-scale DSSR for the anaphora resolution, we also used the Julian. The vocabulary file contained words in all small-scale DSSRs. Since the purpose of the medium-scale DSSR is to capture words in non-order utterances, the accuracy on sentence-level is not always important. However, it needs to handle spontaneous speech utterances. Therefore, the grammar file consisted of the combination of the words in a fixed length; e.g., Noun-PP-Noun-PP-Verb.

4.2 Results

First we evaluated the output selection with a dataset which consists of 20 utterances for each DSSR and 20 out-of-domain utterances such as greetings. The number of test subjects was 10. In other words, we evaluated our method with 1000 utterances: 5 categories (4 DSSRs[1] and LVCSR) $\times$ 20 utterances $\times$ 10 test subjects. The thresh$_{utter}$ and thresh$_{word}$ were 0.26 and 0.08 respectively. These thresholds were determined on a preliminary experiment with another dataset.

The F-value of the output selection was 0.916 on average. In addition, the word recognition accuracy of each DSSR was 0.940 on average. Besides, we verified that the change of the F-value was small even if we changed the thresholds within the compass of 0.20-0.26[2]. Therefore, our method, which was based on the edit distance, for the output selection in a multiple recognizer is simple and robust.

Next, we evaluated the anaphora resolution process in our method combining 3 types of speech recognizers. The dataset of this experiment consisted of 206 utterances that included 53 anaphoric expressions. Figure 5 shows an example of a dialogue in this experiment. In the figure, “###” denotes “change of a speaker” or “change of the location of a robot”. The number of test subjects was 2 persons.

Table 1 shows the experimental result. The baseline in the table denotes our method without the medium-scale DSSR. In other words, the method did not handle the dialogue log DL_1. “Related work” is a scoring based anaphora resolution method which has been proposed by (Shimada *et al.*, 2009). It accumulated the scores of each candidate in the n-previous utterances ($\sum_i^N Score\ k_i$)[3]. To compare the related work with our method fairly, we also applied the medium-scale DSSR to it. The proposed method with the medium-scale DSSR outperformed the baseline, namely a method without the medium-scale DSSR, and the related work, namely another method with the medium-scale DSSR. By using the medium-scale DSSR, the recognition accuracy of words in non-order utterances increased. It led to improvement of the accuracy of the anaphora resolution (64.2 versus 71.7). This result shows the effectiveness of our method incorporating the medium-scale DSSR for the anaphora resolution. The related work was based on the summation of the scores of all candidates in the log. In such method, the existence of noise words, that is insertion errors from the speech recognizer, leads to lower accuracy of the anaphora resolution (68.9 versus 71.7).

Although the accuracy of the anaphora resolution increased by using the medium-scale DSSR, the word accuracy for antecedents was insufficient. Misrecognized words caused the decrease of the anaphora resolution process, especially deletion errors of the speech recognizer. If the outputs of speech recognizers and the resolved anaphoric expressions in previous utterances were completely correct, that is the oracle data, the accuracy of the anaphora resolution became more than 95%. This result shows the significance of the accuracy of speech recognizers that captures the words in non-order utterances. On the other hand, the grammars of our medium-scale DSSR were not considered carefully, that is a simple combination of words without statistical model such

[1] In this evaluation, we did not treat the medium-scale DSSR because it is a recognizer for anaphora resolution.

[2] The best F-value on this experiment was 0.924 in the case that thresh$_{utter}$=0.20.

[3] On the other hand, the proposed method was “$max\ Score\ k_i$”.

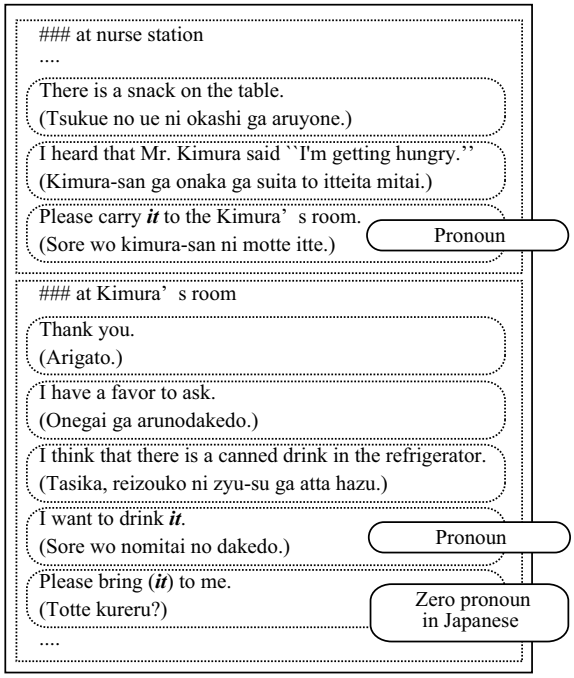

Figure 5: An example of a dialogue for the anaphora resolution.

Table 1: The accuracy of anaphora resolution.

Method	Baseline	Related work	Proposed
Accuracy	64.2%	68.9%	71.7%

as word n-grams. We need to consider the vocabulary and grammar files or the language model of the medium-scale DSSR to improve the accuracy.

In our method, we handle the change of the situation in a dialogue. It is, however, the change of a speaker and the location only. To improve the accuracy of the anaphora resolution, we need to incorporate more detailed situation change model such as topics in the dialogue.

4.3 Related work

For the utterance verification task, many approaches have been proposed. Sako et al. (2006) have reported a method to discriminate a request to a system from a chat using AdaBoost. Machine learning techniques generally need a large amount of training data to generate a classifier with high accuracy. However constructing training data by hand is costly. Isobe et al. (2007) have proposed a multi-domain speech recognition system based on the model likelihoods of the different domain specific language models. The method needs to recalculate a model to select an output. On the other hand, our method only changes two thresholds.

Komatani et. al. (2007) have reported an utterance verification method based on a difference of acoustic likelihood values computed from two recognizers. Kumar et. al. (2005) have utilized Bhattacharyya distance to measure an acoustic similarity of different languages for multilingual speech recognition. Using the difference of acoustic likelihood is adequate for the verification task. Combining a method based on acoustic likelihood with our method is one future work.

For the anaphora resolution task, our method was based on a scoring process using the confidence measure, distance and situation changes. In studies for anaphora resolution on text, machine learning-based methods have been used (Iida *et al.*, 2005; Ng and Cardie, 2002). However ma-

chine learning-based methods need to a large amount of training data. The most famous approach for zero pronouns is the centering theory (Kameyama, 1986). Nariyama (2002) has proposed a method which is an expansion of the centering approach. Minewaki et al. (2005) have reported an utterance interpretation method based on the relevance theory. Incorporating the linguistic knowledge into our method is one of the most effective approaches.

The most critical problem of the anaphora resolution in speech understanding is insertion and deletion errors in dialogue logs, namely the existence of noise words and the lack of the antecedent. Therefore systems need to improve the word recognition accuracy for the anaphora resolution. As a solution for the problem, we applied a medium scale speech recognizer to our method. It is in the category of the ROVER method (Fiscus, 1997). Applying different types of speech recognizers to our method is one future work.

Another approach for the improvement is to repair recognition errors by users. Since our task is an interaction with a robot, repairing errors in a conversation by users is an effective approach. Ogata and Goto (2005) have proposed a speech input interface with a speech-repair function. A dialog processing with visualization of outputs and utterance generation from a robot is one interesting approach in our task.

5 Conclusions

In this paper, we described a speech understanding method based on a multiple speech recognizer. We called it "OGSS model". The method was combination of one LVCSR and several DSSRs. By using this method, we realized a flexible and robust speech understanding method.

In this paper, we evaluated two processes of the method: (1) output selection and (2) anaphora resolution. For the output selection, the method was based on the edit distance between each output. In the experiment, we obtained high F-value (more than 0.9). This result shows that our method is simple and robust. For the anaphora resolution, the method was based on a scoring process of each word with a confidence value in dialogue logs. We also used the distance between an anaphora expression and an antecedent, and change of situation such as speaker's change for the scoring process. Although the proposed method was effective as compared with a baseline, the accuracy was not high (71.7%). The reason why the accuracy of the anaphora resolution was low was that the accuracies of the LVCSR and the medium-scale DSSR were insufficient. To improve the accuracy of the anaphora resolution, we need speech recognizers with more high accuracy for capturing content words in non-order utterances. One approach to solve the problem is to apply a statistical model to the medium-scale DSSR.

Our future work includes (1) a large-scale experiment especially the anaphora resolution, (2) evaluation of the proposed method with other domains and (3) improvement of the accuracy of the medium-scale DSSR.

References

Bouwman, C., J. Sturm, and L. Boves. 1999. Incorporating confidence measures in the Dutch train timetable information system developed in the ARICE project. In *Proceedings of ICASSP*.

Fiscus, J. G. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 347–352.

Iida, R., K. Inui, and Y. Matsumoto. 2005. The issue of combining anaphoricity determination and antecedent identification in anaphora resolution. In *International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE)*, pp. 244–249.

Isobe, T., K. Itou, and K. Takeda. 2007. A likelihood normalization method for the domain selection in the multi-decoder speech recognition system. *IEICE TRANSACTIONS on Information and Systems (Japanese Edition)*, 90(7), 1773–1780.

Kameyama, M. 1986. A property-sharing constraint in centering. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pp. 200–206.

Komatani, K., Y. Fukubayashi, T. Ogata, and H. G. Okuno. 2007. Introducing utterance verification in spoken dialogue system to improve dynamic help generation for novice users. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pp. 202–205.

Komatani, K., R. Hamabe, T. Ogata, and H. G. Okuno. 2005. Generating confirmation to distinguish phonologically confusing word pairs in spoken dialogue systems. In *Proceedings of 4th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pp. 40–45.

Komatani, K. and T. Kawahara. 2000. Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output. In *Proceedings of International Conference on Computational Linguistics (COLING 2000)*, volume 1, pp. 467–473.

Kumar, S. C., V. P. Mohandas, and H. Li. 2005. Multilingual speech recognition: A unified approach. In *Proceedings of InterSpeech 2005*, pp. 3357–3360.

Lee, A., T. Kawahara, and K. Shikano. 2001. Julius - an open source real-time large vocabulary recognition engine. In *Proceedings of Eurospeech*, pp. 1691–1694.

Minewaki, S., K. Shimada, and T. Endo. 2005. Interpretation of utterances based on relevance theory: Toward the formalization of implicature with the maximum relevance. In *Proceedings of the 9th Conference of the Pacific Association for Computational Linguistics (PACLING2005)*, pp. 211–216.

Nariyama, S. 2002. Grammar for ellipsis resolution in japanese. In *In Proceedings of the 9th International conference on Theoretical and Methodological Issues in Machine Translation*, pp. 135–145.

Ng, V. and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 104–111.

Ogata, J. and M. Goto. 2005. Speech repair: Quick error correction just by using selection operation for speech input interfaces. In *Proceedings of Interspeech 2005*, pp. 133–136.

Sako, A., T. Takiguchi, and Y. Ariki. 2006. System request discrimination based on AdaBoost. In *IPSJ technical report. SIG-SLP64*, pp. 19–24.

Shimada, K., A. Uzumaki, M. Kitajima, and T. Endo. 2009. Speech understanding in a multiple recognizer with an anaphora resolution process. In *Proceedings of the 11th Conference of the Pacific Association for Computational Linguistics (PACLING2009)*, pp. 262–267.

Utsuro, T., H. Nishizaki, Y. Kodama, and S. Nakagawa. 2004. Estimating highly confident portions based on agreement among outputs of multiple LVCSR models. *Systems and Computers in Japan*, 35(7), 33–40.