Topic Estimation with Domain Extensibility for Guiding User's Out-of-Grammar Utterances in Multi-Domain Spoken Dialogue Systems

Satoshi Ikeda, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University
Yoshida-Hommachi, Sakyo, Kyoto 606-8501, Japan
{sikeda,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp

Abstract

In a multi-domain spoken dialogue system, a user's utterances are more prone to be out-of-grammar, because this kind of system deals with more tasks than a single-domain system. We defined a topic as a domain about which users want to find more information, and we developed a method of recovering out-of-grammar utterances based on topic estimation, i.e., by providing a help message in the estimated domain. Moreover, the domain extensibility, that is, to facilitate adding new domains, should be inherently retained in multi-domain systems. We therefore collected documents from the Web as training data for topic estimation. Because the data contained not a few noises, we used Latent Semantic Mapping (LSM), which enables robust topic estimation by removing the effect of noise from the data. The experimental results based on using 272 utterances collected with a Woz-like method showed that our method increased the topic estimation accuracy by 23.1 points from the baseline.

Index Terms: multi-domain spoken dialogue system, topic estimation, out-of-grammar utterance

1. Introduction

More and more novices are using spoken dialogue systems through telephones, including cellular phones. They often experience difficulties in using such systems due to speech recognition errors. These kinds of errors are often caused by the *out-of-grammar utterances*, which contain expressions that systems cannot accept because of their limited set of grammar and vocabulary for language-understanding. Out-of-grammar utterances are inevitable because of the trade-off between the building cost and the capability of a language-understanding module. This is an increasingly important issue for multi-domain spoken dialogue systems, because they deal with various tasks. Furthermore, the domains of such systems are developed independently to retain *domain extensibility*, that is, to facilitate adding new domains. In this kind of architecture, the language-understanding module often lacks consistency throughout the whole system, that is, the acceptable expressions differ among domains. This is another reason for out-of-grammar utterances. Therefore, the previously mentioned trade-off and domain extensibility are critical problems that must be solved to improve the quality of multi-domain spoken dialogue systems.

We developed a method for recovering out-of-grammar utterances by estimating users' intentions. We defined a "topic" as a domain about which users want to find more information, and estimated it as the users' intentions. Topic estimation enables the system to provide an appropriate response, such as providing help messages, even for out-of-grammar utterances. Because multi-domain spoken dialogue systems should retain their domain extensibility, we collected documents from the Web to use as training data. Collecting training data from the Web requires less effort than collecting dialogue corpora to use as training data. Because such automatically collected training data contains a significant amount of noise, we used Latent Semantic Mapping (LSM) [1] to remove the effect of noise from the documents in order to have robust topic estimation.

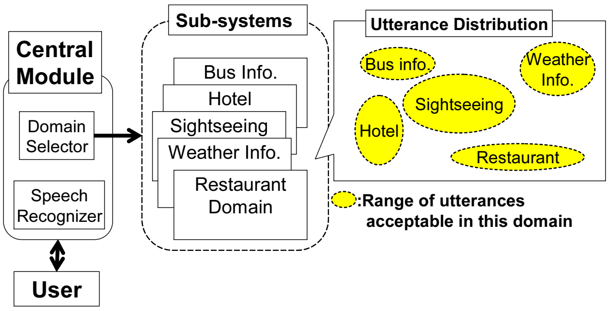

Figure 1: Architecture for multi-domain spoken dialogue system

2. Out-of-grammar Utterances in Multi-domain Spoken Dialogue Systems

2.1. Issues with multi-domain spoken dialogue systems

Domain extensibility is critical for multi-domain spoken dialogue systems, because a large amount of effort is required to develop them. In architecture for such a system, adding and modifying domains should not affect the whole system. It is efficient to reuse the existing domains. Therefore, Lin *et al.* have previously proposed an architecture with domain extensibility [2], which enables system developers to design each domain independently. This architecture is composed of several domain experts that control the dialogues in each domain and a central module that selects an appropriate domain expert to generate a response. The central module does not manage the dialogue to retain domain extensibility. Our multi-domain spoken dialogue system is based on this architecture, and it has five domain experts, as shown in Figure 1.

Out-of-grammar utterances are more critical in multi-domain spoken dialogue systems than in single-domain systems. Multi-domain spoken dialogue systems with domain extensibility have a disadvantage in that users' utterances tend to be out-of-grammar for the following two reasons. The first is that it is difficult for users to guess what expressions the system can accept. In this architecture, each domain accepts a

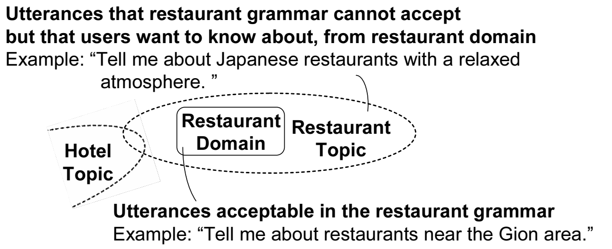
Utterances that restaurant grammar cannot accept
but that users want to know about, from restaurant domain
Example: "Tell me about Japanese restaurants with a relaxed
atmosphere. "
Restaurant Domain
Restaurant Topic
Hotel Topic
Utterances acceptable in the restaurant grammar
Example: "Tell me about restaurants near the Gion area."

Figure 2: Relation between domains and topics

unique kind of expression because for each domain, the grammar for language-understanding is developed independently, so there is inconsistency among the domain grammars throughout the whole system. The second is that the users' utterances often contain various expressions the system cannot handle, because it must deal with various tasks. It is thus difficult for novices to use multi-domain spoken dialogue systems.

2.2. Topic estimation for guiding users' utterances

To deal with out-of-grammar utterances in multi-domain spoken dialogue systems, we estimate a user's intention. We define a *domain* as a sub-system in a multi-domain spoken dialogue system and *in-domain utterances* as utterances that any of the domains will accept. We define *out-of-grammar utterances* as the utterances that none of the domains in the system will accept. We define a *topic* as a domain from which users want to retrieve information, and estimate it as the user's intention. Figure 2 shows the relation between domains and topics.

We aim to build a dialogue system that exhibits the following kind of dialogue management. The topic estimation is executed simultaneously with the conventional grammar-based language-understanding module.

- If the user's utterance is out-of-grammar[1], the topic estimation module is invoked, and the system responds based on the feedback from the module, such as by providing a help message for the estimated domain.
- Otherwise, the system ordinarily selects domains on the basis of feedback from the conventional language-understanding module.

Figure 3 shows an example dialogue that provides help messages based on the topic estimation. In utterance U1, the system cannot understand the user's utterance, because it is out-of-grammar. The system rejects it and estimates its topic. The system understands the user intended to know about the hotel domain, and provides an appropriate help message about the hotel domain as S1.

The topic should be estimated without disrupting the system's domain extensibility. That is, the topic estimation should be applicable even when corpora cannot be sufficiently prepared for a newly added domain. Therefore, we developed a method that does not presuppose training data for each domain.

2.3. Related works

Bohus *et al.* investigated non-understanding errors and 10 recovery strategies for a single-domain spoken dialogue system [4]. Several strategies have previously been investigated, such as the *YouCanSay* strategy, in which the system tells the user what he or she can say at this point in the dialogue. However, these strategies cannot be used in multi-domain spoken dialogue systems, because the current domain for which these strategies should be used needs to be estimated. Lane *et al.* developed a method for detecting utterances whose topics cannot be handled by the system [5]. They estimated topics and detected such utterances with a support vector machine (SVM) or a linear discriminant analysis. However, their method lacks domain extensibility, because they required dialogue corpora collected in advance to be collected in advance, which require a lot of effort.

U1: I'm looking for a place to stay tonight. Rather cheaper ones are preferred. I like a hot spring.
(*Detected as out-of-grammar utterances and rejected. The topic is estimated as "hotel."*)
S1: I do not understand what you say. You can ask about several conditions such as fee, location, leisure around it. For example, you can say "Tell me hotels near the Gion area".
(*A help message for the hotel domain is selected based on the estimated topic.*)
U2: Tell me hotels less than 10,000 yen with a hot spring.
...

Figure 3: Example of dialogue that topic estimation enables

3. Topic Estimation with Domain Extensibility

We estimate topics without disrupting domain extensibility, which is important for multi-domain spoken dialogue systems. We used the following two methods.

1. Collecting training data from the Web.
   This enables topic estimation in situations where corpora about the topic are inadequate, because it is easy to collect training data on any topic from the Web. This allows for topic estimation with domain extensibility.
2. Using LSM to remove the effect of noise from the training data.
   We used LSM because the training data collected from the Web contains documents with other topics as noise. LSM is a technique that is used in natural language processing and applied in areas such as categorizing and summarizing text [6]. LSM can be used to express the conceptual topic of a document as a vector by mapping words and documents onto a low-dimensional space. Therefore, LSM removes the effect of noise from such data and allows for robust topic estimation.

Figure 4 shows an overview of the topic estimation. Our system has six topics: restaurant, sightseeing, bus information, hotel, weather information, and command.

3.1. Collecting training data from the Web

We collected 100,000 sentences from the Web for each five topic, except the command topic. We used a tool for developing language models [7] to collect the documents. We first manually prepared several hundred sentences related to the topic from

[1] Distinction between in-grammar and out-of-grammar utterances will be done by comparing acoustic scores of several ASR results, which has been studied as utterance verification technique [3].

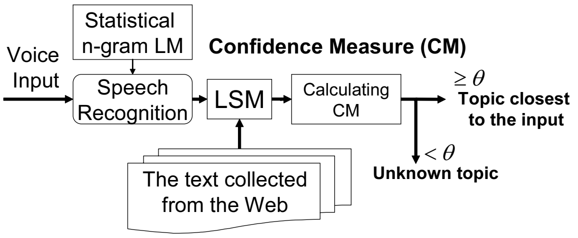

Figure 4: Overview of topic estimation

Wikipedia[2] . We call these texts "seed data". The tool retrieves sentences similar to the seed data from the Web, because the amount of seed data is small. Candidate texts of about 5,000 Web pages were retrieved using about 10 keywords that were selected manually. For example, the keywords related to the hotel topic are "inn" and "stay". The language model was constructed from the seed data and was used to calculate the word perplexity for each sentence collected from the Web. The word perplexity is used to filter the retrieved Web pages, because the retrieved Web texts are not necessarily suitable as training data. The tool selects 100,000 sentences from the retrieved Web texts on the basis of perplexity. We added 10,000 sentences, which were generated by each domain grammar to this training data. For the command topic, which corresponds to the command utterances for the system such as "yes" and "help", we prepared 175 sentences as training data. We randomly divided the training data into d sets ($d = 20$) for each topic, and we made up the training documents. The documents were used to construct a co-occurrence matrix as discussed in the following section.

3.2. Latent semantic mapping

We estimated topics by using LSM [1] to calculate the degree of closeness between a user's utterances and the training documents.

We decomposed the co-occurrence matrix to obtain the k-dimensional vectors of all the training documents. We constructed the ($M \times N$) co-occurrence matrix between the words and the training documents, where M is the vocabulary size, and N is the total number of documents. Here, we denote the number of topics handled by the system as n and the number of training documents for each topic as d. Then, N is represented as $N = n \times d$. We apply the singular value decomposition (SVD) to the matrix and compress its rank to k. The k-dimensional vectors of all the documents are calculated based on the matrix obtained from the SVD. The size of the co-occurrence matrix we constructed is $M = 67533$, $N = 120$, $n = 6$, $d = 20$, and $k = 50$.

Topics are estimated by calculating the cosine distance between the k-dimensional vector of a user's utterance and the k-dimensional vectors of the training documents. First, the user's utterance is recognized using a statistical language model whose vocabulary size is larger than the conventional grammar model used for language-understanding. The k-dimensional vector of the user's utterance is calculated from its ASR result and the matrix obtained from the SVD. The degree of closeness between a topic and a user's utterance is defined as the maximum cosine distance between the k-dimensional vector of a user's utterance and the k-dimensional vectors of d training documents related to the topic. This definition of closeness enables us to express ranges within each topic; the details of this are described in Section 4.3.2.

3.3. Managing unknown-topic utterance

There are some utterances whose topics cannot be estimated directly from themselves. We call them *unknown-topic utterances*. They are classified into two types: *contextual utterances* and *out-of-topic utterances*. The former depends on context, and their topics cannot be estimated uniquely. The latter corresponds to those utterances whose contents are not handled by the system. For example, the utterance, "Tell me the ones less than 5,000 yen" may be a hotel topic or a restaurant topic, which depends on the context of the utterance. The utterance, "Tell me the bank near here" for the hotel and restaurant search system, on the other hand, cannot be handled by the system. We estimate unknown-topic utterances with the following procedure.

First, we obtain T, which is the topic that is most closely related to the user's utterance. We define CM_T as the confidence measure of topic T, which is given by $CM_{t_i} = closeness_{t_i} / \sum_j closeness_{t_j}$, where t_i and t_j are the topics handled by the system, and $closeness_t$ is the degree of closeness between topic t and the user's utterance. If $CM_T > \theta$, the resulting topic is T, where θ is the threshold value. Otherwise, the resulting topic is an unknown topic. When a topic is regarded as unknown, dialogue management and guidance for users are executed based on other information, such as context.

4. Experimental Evaluation

4.1. Dialogue data for evaluation

We evaluated our method by using two kinds of dialogue data. The first kind is the dialogue data used in [8]. This data was collected from 10 subjects by using the 5-domain spoken dialogue system, and it contains 2205 utterances. We removed utterances that have no meaning, like utterance fragments, and obtained 2129 utterances. We call them "with-instruction dialogue data", because they were collected after the subjects had been given instructions. Many in-grammar utterances are contained in this data.

The other kind is the dialogue data we collected from 8 subjects by using a Woz-like method to evaluate more natural utterances. We obtained 272 utterances. We call them "without-instruction dialogue data", because we do not give the subjects examples of the type of utterances that the system can accept. Many utterances in this data are out-of-grammar. They are realistic user utterances similar to those that novice users would speak to the system. All utterances in both kinds of dialogue data were classified manually into seven classes: restaurant, hotel, sightseeing, bus information, weather information, command, and unknown topic.

We used Julius[3] as the speech recognizer, and its language model was constructed from the training data of the LSM. The average word correctness of ASR was 69.6% for with-instruction dialogue data, and 67.3% for without-instruction dialogue data.

4.2. Evaluation of topic estimation

We defined the baseline method as follows:

Baseline method: A topic was estimated using the vector space model based on only the sentences generated by

[2] http://ja.wikipedia.org/

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

Table 1: Correctness of topic estimation for each method

	with-instruction dialogue data	without-instruction dialogue data
baseline	53.5%	37.9%
+ (1) Web collection	51.2%	44.5%
+ (2) LSM	60.8%	45.6%
+ (1) + (2) (our method)	60.4%	61.0%

the domain grammars. Neither the collection of text from the Web (denoted as "Web collection") nor LSM was used.

We compared the correctness of our topic estimation method with those of the baseline method, the method with the Web collection, and the method with LSM. Table 1 lists the correctness of the topic estimation for ASR results of each dialogue data. Here, θ was optimized by trial and error for each condition.

4.3. Discussion

4.3.1. Effectiveness of our method

In the baseline method, the correctness of the topic estimation for the without-instruction dialogue data was much lower than that for the with-instruction dialogue data, as listed in Table 1. This is because the without-instruction dialogue data contained many out-of-grammar utterances. The correctness for this data was increased by 6.6 points from that for the baseline by collecting a large amount of training data from the Web to augment the system's knowledge. On the other hand, the correctness for the with-instruction data decreased by 2.3 points from that for the baseline. This shows that the training data collected from the Web contains a significant amount of noise. The correctness further increased by 16.5 points when we used the LSM, and it was comparable to that for the with-instruction dialogue data. This is because the LSM removed the effect of noise from the training data. These indicate that both collecting training data from the Web and using LSM are essential for our topic estimation method and that our method is robust against out-of-grammar utterances. For example, even for utterances, such as "Tell me recommended visitor attraction," and "Tell me the ones with large room," which do not have content words, our method correctly estimated their topics as "sightseeing topic" and "hotel topic".

4.3.2. Range of topics

We calculated the average cosine distance between the center of a topic and training documents related to the topic, as listed in Table 2. The center of a topic is defined as the average vector of d training documents related to the topic. As the average cosine distance decreases, the range of the topic broadens.

On the basis of the results listed in Table 2, our method represents ranges within each topic successfully. Command, weather information, and bus information have higher values. This means that there are relatively fewer variations in the language expressions for these topics. In fact, the bus topic has little variation in the user's utterances, because it deals with only bus information. On the other hand, the hotel and sightseeing topics have lower values. The sightseeing topic has a large variation in the user's utterances, because it deals with various tasks, such as "searching temples," "zoo," "museum."

Table 2: Average cosine distance between center of topic and training documents related to the topic

topic	average cosine distance
restaurant	0.93
hotel	0.84
sightseeing	0.84
bus information	0.93
weather information	0.90
command	0.99

5. Conclusion

We developed a method for estimating topics with domain extensibility in multi-domain spoken dialogue systems. A topic is defined as a domain about which a user wants to find more information, and it was used to augment the type of language expressions that the system can handle. This was possible by collecting training data from the Web and by using LSM. Consequently, appropriate responses such as a help message can be generated for an estimated domain even for expressions that could not previously be handled by the system.

The experimental results for the "without-instruction" dialogue data showed that our method can be used to estimate topics with 23.1 points higher accuracy than the baseline method. This shows that our method can be used to robustly estimate topics even from out-of-grammar utterances.

The topic estimation we reported in this paper used only information obtained from the current utterance. By introducing the contextual information we have previously developed [8], the topic estimation and dialogue management based on it will be more accurate. We will include this in our future work.

Acknowledgements: We are grateful to Mr. Misu and Prof. Kawahara of Kyoto University for allowing us to use the document-collecting program they developed [7].

6. References

[1] J. R. Bellegarda, "Latent semantic mapping." *IEEE Signal Processing Mag.*, vol. 22, no. 5, pp. 70–80, 2005.

[2] B. Lin *et al.*, "A distributed agent architecture for intelligent multi-domain spoken dialogue systems," in *Proc. ASRU*, 1999.

[3] N. Kitaoka *et al.*, "Confidence measure and rejection based on correctness probability of rejection candidate," *Systems and Computers in Japan*, vol. 35, no. 11, pp. 91–102, 2004.

[4] D. Bohus and A. I. Rudnicky, "Sorry, I didn't catch that! - an investigation of non-understanding errors and recovery strategies," in *Proc. SIGDial*, 2005, pp. 128–143.

[5] I. R. Lane *et al.*, "Topic classification and verification modeling for out-of-domain utterance detection," in *Proc. ICSLP*, 2004, pp. 2197–2200.

[6] J. Steinberger and K. Ježek, "Using latent semantic analysis in text summarization and summary evaluation," in *Proc. ISIM*, 2004, pp. 93–100.

[7] T. Misu and T. Kawahara, "A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts," in *Proc. Interspeech*, 2006, pp. 9–12.

[8] K. Komatani *et al.*, "Multi-domain spoken dialogue system with extensibility and robustness against speech recognition errors," in *Proc. SIGDial*, July 2006, pp. 9–17.