Extensibility Verification of Robust Domain Selection against Out-of-Grammar Utterances in Multi-Domain Spoken Dialogue System

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

We developed a robust domain selection method and verified its extensibility. An issue in domain selection is its robustness against out-of-grammar utterances. It is essential to generate correct system responses because such utterances often cause domain selection errors. We therefore integrated the topic estimation results and the dialogue history to construct a robust domain classifier. Another issue is that domain selection should be performed within an extensible framework, because the system is often modified and extended. That is, the classifier should still have high performance without reconstructing it after adding new domains. The extensibility of our method was not experimentally verified yet, because it requires a lot of effort to collect new dialogue data after extending the system. Therefore, we verified extensibility without collecting new data. We constructed the classifier by leaving out some domains in the dialogue data and then evaluated its accuracy as the classifier for the data where the left-out domains were virtually added.

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

1. Introduction

Multi-domain spoken dialogue systems deal with various tasks, such as searching for restaurants and retrieving bus information. Such systems are convenient for users, although a large amount of effort is required to develop them. System developers need to modify or add tasks to the existing system to handle users' various requests. **Extensibility**, that is, the ability to modify and add tasks that a system deals with, is an inevitable element in such systems. To retain extensibility, these systems are developed by integrating single-domain systems that handle each task [1]. Therefore, **domain selection**, i.e., determining which subsystem in the multi-domain systems should respond to a user's request, is essential for such systems. Our multi-domain spoken dialogue system is based on this architecture, and, as shown in Figure 1, it consists of five domains such as restaurant, hotel, sightseeing, bus information, and weather information.

An issue in domain selection is its robustness against out-of-grammar utterances. It is essential to generate correct system responses because such utterances caused domain selection errors in our previous method [2] and conventional methods [1, 3]. Therefore, we develop a robust domain selection against out-of-grammar utterances by integrating topic estimation results [4] and the dialogue history [2]. Because topic estimation is robust against out-of-grammar utterances while it does not take the dialogue history into consideration, we integrated these two. In this paper, we evaluated the robustness of our method.

Another issue is that domain selection should be performed within an extensible framework. That is, when we add new domains after the domain classifier is once constructed, the classifier should still have high performance without reconstructing it by using data collected with the new domains. Our method was constructed in an extensible framework [2, 5]. This framework is reusable when the system is extended and features used for the classification are easily obtained even for newly added domains. The problem is that it was not experimentally verified. This is because experimental verification requires a lot of effort to collect new dialogue data after extending the system. In this paper, we verify extensibility by evaluating the domain selection accuracy when domains were virtually added, instead of collecting new dialogue data. That is, we constructed the classifier by leaving out some domains in the dialogue data and then evaluated its accuracy as the classifier for five-domain systems, where all collected data was used.

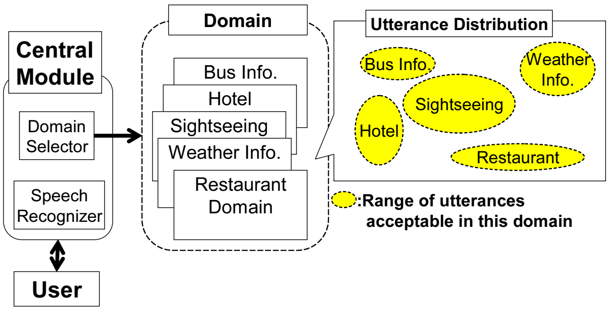

Figure 1: Architecture of our system

2. Robust domain selection using dialogue history and topic estimation

We integrated the topic estimation result and the dialogue history. This integration enables robust domain selection because the topic estimation and use of dialogue history have complementary information. The topic estimation uses only information obtained from a single utterance while dialogue history takes the context into consideration. On the other hand, the dialogue history is often impeded by out-of-grammar utterances while topic estimation results are relatively more reliable for them. The overview of our domain selection is shown in Figure 2. Our domain selection consists of the following two parts. More details were presented in our other paper [5].

2.1. Topic estimation for dealing with out-of-grammar utterance

We first define a *topic* as a domain from which the user wants to retrieve information, and estimate it as the user's intention. We

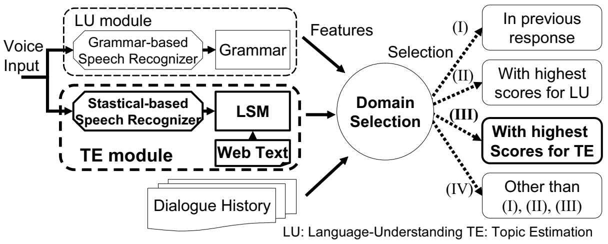
Voice Input
LU module
Grammar-based Speech Recognizer
Grammar
Stastical-based Speech Recognizer
LSM
Web Text
TE module
Dialogue History
Features
Selection
Domain Selection
(I)
In previous response
(II)
With highest scores for LU
(III)
With highest Scores for TE
(IV)
Other than (I), (II), (III)
LU: Language-Understanding TE: Topic Estimation

Figure 2: Overview of domain selection

Table 1: Features used in our previous paper [2]

P1:	number of affirmatives after entering the domain
P2:	number of negations after entering the domain
P3:	whether the domain appeared before
P4:	number of changed slots after selecting (I)
P5:	number of changed slots after entering the domain
P6:	ratio of changed slots
P7:	ratio of user's negative answers in the domain
P8:	posteriori probability of N-best candidates of ASR results for LU interpreted in (I)
P9:	average of words' confidence score for the best candidate in (I)
P10:	posteriori probability of N-best candidates of ASR results for LU interpreted in (Ⅱ)
P11:	average of words' confidence score for the best candidate of ASR results for LU in (Ⅱ)
P12:	whether the language-understanding result is negative after selecting (I)
P13:	dialogue state after selecting (Ⅱ)
P14:	whether (Ⅱ) has appeared before

estimated topics by calculating the closeness between the user's utterance and the training data collected from the Web by using Latent Semantic Mapping (LSM) [4]. The topic estimation module in Figure 2 shows a brief overview of the topic estimation process. Our system has five topics corresponding to each domain and the command topic, which corresponds to the command utterances for the system such as "yes" and "undo." We first collected Web texts for each topic by using a tool for developing language models [6], and merged sentences generated by each domain grammar. The ASR for the topic estimation uses a statistical language model trained with the collected Web texts. We then used LSM [7] because the training data collected from the Web contain documents with other topics as noise. It removes the effect of noise from such data and allows for the robust topic estimation.

2.2. Integrating dialogue history and topic estimation

We design domain selection as the following four-class categorization: (I) the previous domain, (Ⅱ) the domain with the highest score for language-understanding, (Ⅲ) the domain with the highest score for topic estimation, and (Ⅳ) other domains. We constructed a domain classifier using machine learning. Here, we describe the features used to construct it.

In addition to the information listed in Table 1, which was used in our previous work [2], we adopted new features, as listed in Table 2. These features are obtained from the topic estimation result. Using this information enables the system to select correct domains even for out-of-grammar utterances.

Table 2: Features of topic estimation result

T1:	closeness between (Ⅲ) and ASR result for TE
T2:	confidence measure of (Ⅱ)
T3:	difference of closeness to ASR result for TE between (Ⅱ) and (Ⅲ)
T4:	difference of confidence measures between (Ⅱ) and (Ⅲ)
T5:	whether (Ⅲ) is the same as (Ⅱ)
T6:	whether (Ⅲ) is the same as (I)
T7:	whether (Ⅲ) is the same as "command"
T8:	duration of ASR result for TE (number of phoneme in recognition result)
T9:	acoustic score of ASR result for TE
T10:	difference of acoustic scores per phoneme between candidates selected as (Ⅲ) and (I)

TE: Topic Estimation

The features from T1 to T4 are defined so that they represent the confidence in the topic estimation result. We defined the confidence measure of topic T used in T2 and T4 as $CM_T = closeness_T - \frac{1}{N}\sum_j closeness_{t_j}$, where T and t_j are topics handled by the system, $closeness_t$ is the degree of closeness between topic t and the user's utterance, and N is the number of topics. We also adopted the features from T5 to T7 to represent the relation between (I), (Ⅱ) and (Ⅲ). For example, if the topic estimation result is the same as (I), the system prefers (I). We defined feature T8 because an utterance whose duration is too short often causes errors in the estimation of the topic. The features T9 and T10 represent whether the user's utterance is out-of-grammar. If the user's utterance seems so, the system does not prefer (Ⅱ).

Note that the features listed here were selected by backward stepwise selection, in which a feature survives if the performance in the domain classification is degraded when it is removed from a feature set one by one. We had originally prepared 43 features for the initial set.

2.3. Example dialogue

Our robust domain selection enables the system to generate concrete responses even for out-of-grammar utterances, as shown in Figure 3. In utterance U2, the system cannot understand the user's utterance, because it is out-of-grammar. The system does not accept the language-understanding results for U2, and provides help messages based on the domain (in this case, restaurant) derived from the topic estimation result in S2.

3. Extensible architecture and its verification

3.1. Extensibility of our domain selection

Our domain selection is applied within an extensible framework. To retain extensibility, domain selection should fulfill the following two requirements:

1. Features related to newly added domains can be obtained.
2. The domain selection framework is reusable

If a lot of effort is required to obtain features related to new domains, the domain selection cannot be applied immediately to the extended system. Similarly, if the classifier is not reusable

Table 3: Confusion matrix in our domain selection

Reference label \ Output	(I)	(II)	(III)	(IV)	Total (recall)
(I): in previous response	1345	30	31	36	1442 (93.3%)
(II): with highest score for LU	85	256 + 17†	13	9	380 (67.4%)
(III): with highest score for TE	55	11	51	3	120 (42.5%)
(IV): others	138	14	14	83	249 (33.3%)
Total (precision)	1623 (82.9%)	328 (78.0%)	109 (46.8%)	131 (63.4%)	2191 (79.2%)

LU: Language Understanding TE: Topic Estimation

†: These include 17 errors because of random selection when there were several domains having the same highest scores.

U1: **Tell me the address of Holiday Inn.**

S1: The address of Holiday Inn is ... (Selected domain: hotel)

U2: **I want Tamanohikari near there.** (Correct domain: restaurant)
Tamanohikari (name of liquor) is an out-of-vocabulary word, and is misrecognized as a spot of Tamba-bashi (name of place). (Domain: sightseeing*)*

S2: I do not understand what you said. You can ask about several conditions such as location, price and food type about restaurants. For example, you can say "Tell me restaurants near the Gion area". (Selected domain: restaurant)

Figure 3: Example dialogue including out-of-grammar utterances

for the extended system, we have to collect new dialogue data and reconstruct the classifier. Our domain selection method fulfills these two requirements as described below.

First, we developed a topic estimation method that is easily applied to newly added domains. The texts as training data are automatically collected from the Web to reduce the effort required to collect it. This approach makes it possible to easily obtain the features of topic estimation results related to the newly added domain. The conventional method of topic estimation [8] disrupts extensibility because it requires the dialogue data collected by hand. Features listed in Table 1 are also designed as facilitating to obtain from newly added domain. Therefore, all of the features used in our method are easily obtained when domains are added.

Second, our domain selection framework of four-class categorization, which is described in Section 2.2, does not change when the number of domains changes. This is because the categorization is defined as relative choices between previous and current domains, such as "the same as the previous domain". That is, no concrete domain name such as restaurant is used in the classifier. The used features listed in Tables 1 and 2 are also independent of specific domain, accordingly. This framework therefore enables us to avoid reconstructing the classifier when the number of domain changes, and it can be reused when the system is extended.

3.2. Extensibility verification without collecting new data

We verified the extensibility of our domain selection without collecting new dialogue data. The problem of extensibility verification is that it requires a lot of effort to collect new dialogue data. We calculated domain selection accuracies for the cases when domains are virtually added. That is, we leave out some domains when constructing the classifier and then add the left-out domains into the evaluation data. This verification method does not need to collect new dialogue data and can be applied by using dialogue data we have already collected.

The concrete procedure to verify the extensibility is as follows. The removed domain is denoted as D. We first obtained features from the topic estimation module after removing training data for D. The classifier was then trained after removing utterances whose label for either (I) or (II) was D. By using the same classifier, which was constructed with no information related to D, we calculated the classification accuracies for data of the original five-domain system. Note that the topic estimation model and the features listed in Tables 1 and 2 were recalculated for the five-domain data. This can be easily obtained without disrupting extensibility, as explained in Section 3.1.

4. Experimental evaluation

4.1. Dialogue data for evaluation

We evaluated our method by using dialogue data collected from 10 subjects [2]. It contains 2191 utterances. This data was collected by using the following procedure. First, to get accustomed to the timing to speak, the subjects used the system by following a sample scenario. They then used the system by following three scenarios, where at least three domains were contained.

We used Julian[1], a grammar-based speech recognizer, for the language understanding. Its grammar rules correspond to those used in the language-understanding modules in each domain. We also used Julius, a statistical-based speech recognizer, to estimate the topic. Its language model was constructed from the training data collected for the topic estimation. A 3000-state Phonetic Tied-Mixture (PTM) model was used as an acoustic model. The ASR accuracies were 63.3% and 69.6% for each.

Accuracies for domain selection were calculated per utterance. When there were several domains that had the same score after domain selection, one domain was randomly selected from them. We used C5.0 [9] as a classifier. The performance was calculated with a 10-fold cross validation. Reference labels of the domain selection were given by hand for each utterance from (I), (II), (III) or (IV) described in Section 2.2. Note that the labels are given priority in this order when multiple labels can be given for an utterance.

4.2. Evaluation of robustness

We first calculated the accuracy of domain selection with our method. Table 3 lists the classification results in our method

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

Table 4: Accuracy when 4 domains are used to construct the classifier

Domain used for training					Accuracy
restaurant	hotel	sightseeing	bus	weather	(# error)
	○	○	○	○	75.4% (538)
○		○	○	○	75.9% (527)
○	○		○	○	76.1% (524)
○	○	○		○	75.6% (534)
○	○	○	○		75.9% (528)
Average accuracy of our method					75.8%
Baseline					71.2%

as a confusion matrix. The accuracy of our domain selection is 79.2% when the classifier is trained using five domains. Our method can select the correct domain for utterances that the conventional method cannot select the correct domain by detecting (Ⅲ). In fact, using our method enabled the system to successfully select correct domains for 51 of 120 utterances of (Ⅲ). Our method can also avoid successive domain selection errors by detecting (Ⅳ). Our method could select the correct domain for 83 of 249 utterances of (Ⅳ). These are the utterances with an unreliable dialogue history. In fact, 44 domain selection results of their previous utterances were incorrect out of these 83 utterances correctly classified as (Ⅳ).

4.3. Evaluation of extensibility

We evaluated the extensibility of our method by comparing it with a baseline method that does not use machine learning. The baseline method and our method are described below:

Baseline method: A domain having an interpretation with the highest score in the N-best candidates of the speech recognition was selected, after adding α for the acoustic likelihood of the speech recognizer if the domain was the same as the previous one. We set α with which the domain selection accuracy was maximized. This baseline method is equivalent to the conventional method [1].

Our method: Domains were selected by using the classifier trained with three- or four-domain data, instead of the one trained with five-domain data, as described in Section 3.2.

In the baseline method, the smallest number of domain selection errors was 630 when $\alpha = 35$, and the accuracy was 71.2% (= 1561/2191). In our method, the accuracy when four domains and three domains were used to construct the classifier is respectively listed in Tables 4 and 5. The upper limit of our method is 79.2%, as shown in Table 3. These tables show that domain selection errors increase as fewer domains are used for training, but the average accuracy of our method is higher than that of the baseline. The average error reduction rate was 15.8% and 9.3% for each. This result shows our method is not greatly affected by the number of domains and has extensibility. This is because our domain selection framework is based on the classification of relative choices among domains, and because features independent of specific domains are used.

5. Conclusion

We developed a robust domain selection method and verified its extensibility without collecting new dialogue data. The issue in robustness is that out-of-grammar utterances often cause domain selection errors in our previous method [2]. Therefore, we first integrated the topic estimation result and the dialogue history, and constructed robust domain classifier. This integration enables robust domain selection because the topic estimation results and the dialogue history have complementary information.

Table 5: Accuracy when 3 domains are used to construct the classifier

Domain used for training					Accuracy
restaurant	hotel	sightseeing	bus	weather	(# error)
		○	○	○	74.5% (558)
	○		○	○	70.3% (649)
	○	○		○	73.3% (584)
	○	○	○		74.3% (564)
○			○	○	74.2% (565)
○		○		○	74.4% (561)
○		○	○		74.9% (551)
○	○			○	75.0% (547)
○	○		○		74.7% (554)
○	○	○			73.4% (582)
Average accuracy of our method					73.9%
Baseline					71.2%

We then verified extensibility of our method. The issue in extensibility is that it is not experimentally verified because a large amount of effort is required to collect new dialogue data. Therefore, we constructed the classifier by leaving out some domains and then evaluated its accuracy as the five-domain classifier, where all data was used.

In this paper, the domain selection accuracy is verified when domains were virtually added. Thus, we need to evaluate our method when domains are really added, although it needs a lot of effort. We also should compare the results from our verification and those by using real dialogue data to verify the validity of our method. We will include these in our future work.

Acknowledgements We are grateful to Dr. Mikio Nakano of Honda Research Institute Japan and Mr. Naoyuki Kanda for their cooperation in developing the multi-domain system used to collect the evaluation data.

6. References

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

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

[3] I. O’Neill *et al.*, “Cross domain dialogue modelling: an object-based approach,” in *Proc. ICSLP*, 2004, pp. 205–208.

[4] S. Ikeda *et al.*, “Topic estimation with domain extensibility for guiding user’s out-of-grammar utterance in multi-domain spoken dialogue systems,” in *Proc. Interspeech*, 2007, pp. 2561–2564.

[5] S. Ikeda *et al.*, “Integrating topic estimation and dialogue history for domain selection in multi-domain spoken dialogue systems,” in *Proc. IEA/AIE*, 2008, pp. 294–304.

[6] T. Misu and T. Kawahara, “A bootstrapping approach for developing language model of new spoken dialogue systems by selecting Web texts,” in *Proc. ICSLP*, 2006, pp. 9–12.

[7] J. Bellegarda, “Latent semantic mapping.” *IEEE Signal Processing Mag.*, vol. 22, no. 5, pp. 70–80, 2005.

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

[9] J. R. Quinlan, *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann, 1993, http://www.rulequest.com/see5-info.html.