A Two-Stage Domain Selection Framework for Extensible Multi-Domain Spoken Dialogue Systems

Mikio Nakano
Honda Research Institute Japan
Wako, Saitama, Japan
nakano@jp.honda-ri.com

Shun Sato
Tokyo Denki University
Hatoyama, Saitama, Japan
rela.relakuma@gmail.com

Kazunori Komatani
Nagoya University
Nagoya, Aichi, Japan
komatani@nuee.nagoya-u.ac.jp

Kyoko Matsuyama*
Kyoto University
Kyoto, Kyoto, Japan
matuyama@kuis.kyoto-u.ac.jp

Kotaro Funakoshi
Honda Research Institute Japan
Wako, Saitama, Japan
funakoshi@jp.honda-ri.com

Hiroshi G. Okuno
Kyoto University
Kyoto, Kyoto, Japan
okuno@i.kyoto-u.ac.jp

Abstract

This paper describes a general and effective domain selection framework for multi-domain spoken dialogue systems that employ distributed domain experts. The framework consists of two processes: deciding if the current domain continues and estimating the probabilities for selecting other domains. If the current domain does not continue, the domain with the highest activation probability is selected. Since those processes for each domain expert can be designed independently from other experts and can use a large variety of information, the framework achieves both extensibility and robustness against speech recognition errors. The results of an experiment using a corpus of dialogues between humans and a multi-domain dialogue system demonstrate the viability of the proposed framework.

1 Introduction

As spoken dialogue interfaces are becoming more widely utilized, they will be expected to be able to engage in dialogues in a wide variety of topics. Particularly, spoken dialogue interfaces for office robots (Asoh et al., 1999) and multimodal kiosk systems (Gustafson and Bell, 2000) are expected to deal with people's various requests, unlike automated call center systems that are dedicated to specific tasks.

One effective methodology to build such a system is to integrate systems in small domains by employing *distributed multi-domain system architecture*. This architecture has distributed modules that independently manage their own dialogue state and knowledge for speech understanding and utterance generation (e.g., Lin et al. (1999)). From an engineering viewpoint, such architecture has an advantage in that each domain expert can be designed independently and that it is easy to add new domains. It enables each domain expert to employ a dialogue strategy very different from those for other domains. For example, the strategy may be frame-based mixed-initiative, finite-state-based system-initiative, or plan-based dialogue management (McTear, 2004).

One of the crucial issues with distributed multi-domain spoken dialogue systems is how to select an appropriate domain for each user utterance so that the system can appropriately understand it and answer it. So far several methods have been proposed but none of them satisfy two basic requirements at the same time: the ability to be used with a variety of domain experts (**extensibility**) and being robust against ASR (Automatic Speech Recognition) errors (**robustness**). We suspect that this is one of the main reasons why not many multi-domain spoken dialogue systems have been developed even though their utility is widely recognized.

This paper presents a new general framework for domain selection that satisfies the above two requirements. In our framework, each expert needs to have two additional submodules: one for estimating the probability that it is newly activated, and one for deciding domain continuation when it is already activated. Since these submodules can be designed independently from those of other experts, there is no restriction on designing experts in our framework,

*Currently with Panasonic Corporation.

Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 18–29,
Portland, Oregon, June 17-18, 2011.

and thus extensibility is achieved. Robustness is also achieved because those submodules can be designed so that they can utilize domain-dependent information, including information on speech understanding and dialogue history, without detracting from extensibility. Especially the submodule for deciding domain continuation has the ability to utilize dialogue history to avoid erroneous domain shifts that often occur in previous approaches. Note that we do not focus on classifying each utterance without contextual information (e.g., Chu-Carroll and Carpenter (1999)). Rather, we try to estimate the user intention with regard to continuing and shifting domains in the course of dialogues.

In what follows, Section 2 explains the distributed multi-domain spoken dialogue system architecture and requirements for domain selection. Section 3 discusses previous work, and Section 4 presents our proposed framework. Section 5 describes an example implementation and its evaluation results, and Section 6 concludes the paper.

2 Domain Selection in Multi-Domain Spoken Dialogue Systems

2.1 Distributed Architecture

In distributed multi-domain spoken dialogue architecture (Figure 1), distributed modules independently manage their own dialogue state and knowledge for speech understanding and utterance generation (Lin et al., 1999; Salonen et al., 2004; Pakucs, 2003; Nakano et al., 2008). Although those modules are referred to with various names in that literature, we call them *domain experts* in this paper. In this architecture, when an input utterance is received, its ASR results are sent to domain experts. They try to understand the ASR results using their own knowledge for understanding. The domain selector gathers information from those experts and decides which expert should deal with the utterance and then decide on the system utterances. In this paper, the domain expert engaging in understanding user utterances and deciding system utterances is called *activated*.

2.2 Example Systems

So far many multi-domain spoken dialogue systems based on distributed architecture have been

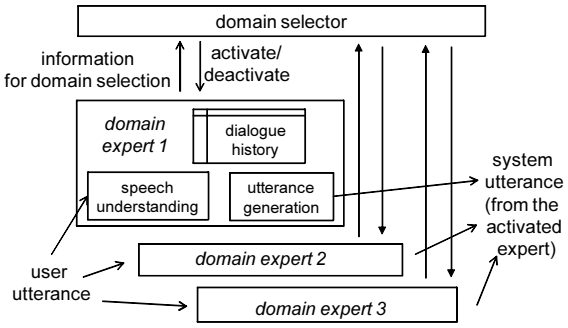

Figure 1: Distributed multi-domain spoken dialogue system architecture.

built and have demonstrated their ability to engage in dialogues in a variety of domains. For example, several systems integrated information providing and database searches in multiple domains (Lin et al., 1999; Komatani et al., 2006; O'Neill et al., 2004; Gustafson and Bell, 2000). Some other systems integrated domain experts that employ very different dialogue strategies. Lee et al. (2009) and Nakano et al. (2006) integrated task-oriented and non-task-oriented dialogue managements. Nakano et al. (2008) integrated domain experts for not only dialogues but also tasks requiring physical actions.

Below we explain an example system that we used to collect dialogue data for the domain selection experiment described in Section 5. It is a Japanese system that can provide information on UNESCO World Heritage Sites using speech, slides, and Microsoft Agent[1] gestures. It employs the following ten domain experts:

A *question answering (QA) expert*: It has a database consisting of question-answer pairs on World Heritage Sites (Narimatsu et al., 2010). Each question-answer pair consists of a set of example question sentences and answers to them. The answers consist of one or more utterances, plus slides. Keyphrase-based matching is used to select an example question that matches the input utterance.

Eight *interactive presentation (IP) experts*: Each of them explains in detail a famous World Heritage Site and adds a detailed explanation or skips some explanation according to the user's interruptions (Nishimura et al., 2007). Patterns of interruption utterances are described in the expert's knowledge base.

[1] http://www.microsoft.com/MSAgent/

	utterance	expert
S1:	Hello. I can answer questions about famous World Heritage Sites in Italy, Germany, Greece, and other areas. I can also explain some World Heritage Sites in detail. Please ask me anything.	QA
U1:	Show me the list of the World Heritage Sites in Italy.	QA
S2:	(show a slide of the list) I can tell you about the World Heritage Sites in these areas in Italy.	QA
U2:	Can you show me the list of World Heritage Sites you can explain in detail?	QA
S3:	(show a slide of the list) I can explain these in detail.	QA
U3:	Can you explain Maritime Greenwich?	RU
S4:	Maritime Greenwich, right?	RU
U4:	Yes.	RU
S5:	(show a slide on Greenwich) Okay. I'll explain Maritime Greenwich, a historic port town known for Greenwich Mean Time.	IP1
U5:	What is Greenwich Mean Time?	IP1
S6:	Greenwich Mean Time is the mean solar time at the Royal Greenwich Observatory.	IP1

S means system utterance and U means user utterance. IP1 is one of the IP experts.

Figure 2: An example dialogue with the system for data collection (translation from Japanese).

A *request understanding (RU) expert*: It understands the user's request to start one of the interactive presentations explained above, and engages in a dialogue to confirm the request. When the understanding finishes, the understood request is sent to a module called *task planner* (Nakano et al., 2008; Nakano et al., 2011). The task planner then activates another expert to perform the requested presentation (S5 in Figure 2).

Figure 2 shows an example dialogue between a human and this system. Note that user utterances are relatively short and include words related to specific World Heritage Sites or area names. If those words are misrecognized, domain selection is difficult unless dialogue context information is used.

This figure also indicates the domain experts that understood each user utterance and selected each system utterance. The domain expert that should deal with a user utterance is decided based on the set of user utterances that the expert is designed to deal with. The domains of utterances U1 and U3 are different because the QA expert has knowledge for understanding U1 and the RU expert has knowledge for understanding U3. Thus, in this study, the domain of each utterance is determined based on the design of the experts employed in the system. If none of the experts can deal with an utterance, it is considered as an out-of-domain utterance. Sometimes the correct domain needs to be determined using contextual information. For example, utterance U4 "Yes" can appear in all domains, but, since this is a reply to S4, its domain is RU.

This definition of domain is different from that of domain (or topic) recognition and adaptation studies in text, monologue, and human-human conversation processing, in which reference domains are annotated based on human perspectives rather than system perspectives. From a human perspective, all user utterances in Figure 2 may be in "World Heritage Site" domain. However, it is not always easy to build domain experts according to such domain definitions, because different dialogue tasks in one such domain may require different dialogue strategies (such as question answering and request understanding).

2.3 Requirements for Domain Selection

We pursue a method for domain selection that can be used in distributed architecture. Such a method must satisfy the following two requirements.

Extensibility It must not detract from the extensibility of distributed architecture, that is, any kind of expert must be able to be incorporated, and each expert must be able to be designed independently from other experts. This requires the interface between each domain expert and the domain selector to be as simple as possible.

Robustness It needs to be robust against ASR errors; that is, the system needs to be able to avoid erroneous domain transition caused by ASR errors.

3 Previous Work

So far various methods for domain selection have been proposed, but, as far as we know, no method satisfies both extensibility and robustness. Isobe et al. (2003) estimate a score for each domain from the

ASR result and select the domain with the highest score (hereafter referred to as RECSCORE). Since each domain expert has only to output a numeric score, it satisfies extensibility. However, because this method does not take into account dialogue context, it tends to erroneously shift domains when the score of some experts becomes high by chance. For example, if U4:"Yes" in Figure 2 is recognized as "Italy" with a high recognition score in the QA expert, the domain erroneously shifts to QA and the system explains about World Heritage Sites in Italy. Thus this method is not robust.

To avoid erroneous domain shifts, Lin et al. (1999) give preference to the *preceding domain* (the domain in which the previous system utterance was made) by adding a certain value to the score of the preceding domain (hereafter called RECSCORE+BIAS). However, to what extent the domain tends to continue varies depending on the dialogue context. For example, if a dialogue task in one domain finishes (e.g., when an IP expert finishes its presentation and says "This is the end of the presentation. Do you have any questions?"), the domain is likely to shift. So, adding a fixed score does not always work. O'Neill et al.'s (2004) system does not change the dialogue domain until it finishes a task in the domain, but it cannot recover from erroneous domain shifts.

To achieve robustness against ASR errors, several domain selection methods based on a classifier that uses features concerning dialogue history as well as ones concerning speech understanding results have been developed (Komatani et al., 2006; Ikeda et al., 2008; Lee et al., 2009). These studies, however, use some features available only in some specific type of domain experts, such as features concerning slot-filling, so they cannot be used with other kinds of domain experts. That is, these methods do not satisfy extensibility.

Methods that use classifiers based on word (and n-gram) frequencies have been developed for utterance classification (e.g., Chu-Carroll and Carpenter (1999)), topic estimation for ASR of speech corpora (e.g., Hsu and Glass (2006) and Heidel and Lee (2007)) and human-human dialogues (Lane and Kawahara, 2005). These methods can be applied to domain selection in multi-domain spoken dialogue systems. However, since they require training data in the same set of domains as the target system, it detracts from extensibility. In addition, they are not robust because they cannot utilize a variety of dialogue and understanding related features. Word frequencies are not always effective when two domains share words as in our system described in Section 2.2.

4 Proposed Framework

4.1 Basic Idea

To achieve extensibility, we need to restrict the information that each expert sends to the domain selector to a simple one such as numeric scores. Although RECSCORE and RECSCORE+BIAS satisfy this, they would not achieve high accuracy as explained above.

One possible extension to those methods to improve accuracy is to use not only recognition scores but also various expert-dependent features such as ones concerning dialogue history and speech understanding. Each expert first estimates the probability that the input utterance is in its domain using such features, and then the expert with the highest probability is selected (hereafter called MAXPROB). This method retains extensibility because the domain selector does not directly use those expert-dependent features. However, it suffers from the same problem as RECSCORE and RECSCORE+BIAS; if one of the experts other than the preceding domain's expert outputs a high probability by mistake, the domain shifts regardless of the dialogue state in the preceding domain's expert.

We focus attention on the fact that the domain does not often shift. Our idea is to decide if the domain continues or not by using information available in the preceding domain's expert. This prevents erroneous domain shifts when the utterance is considered not to change the domain. When it is decided that the currently active domain does not continue, each remaining expert estimates the probability of being newly activated using information available in the expert, and the expert whose probability is the highest is selected as the new domain expert.

We further refine this idea in two ways. One is by taking into account how likely the input utterance is to activate one of the other domain experts. We propose to use the maximum value of probabilities for other experts' activation (*maximum activation prob-*

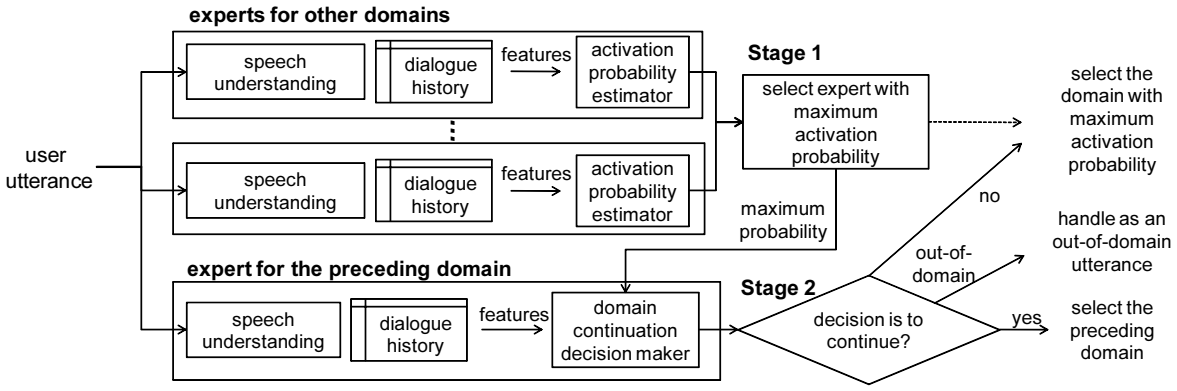

Figure 3: Two-stage domain selection framework.

ability) in the decision regarding domain continuation. Since the maximum activation probability is just a numeric score, this does not spoil extensibility. Unlike RECSCORE and RECSCORE+BIAS, in our method, even if the maximum activation probability is very high, the preceding domain's expert can decide to continue or not to continue based on its internal state. This makes it possible to retain robustness.

The other refinement is to explicitly deal with utterances that are not in any domains (*out-of-domain (OOD) utterances*). They include fillers and murmurs. They should be treated separately, because they appear context-independently. So we make the expert detect OOD utterances when deciding domain continuation. That is, it performs three-fold classification, *continue*, *not-continue*, and *OOD*.

4.2 Two-Stage Domain Selection Framework

This idea can be summarized as a domain selection framework which consists of two stages (Figure 3). It assumes that each domain expert has two submodules: *activation probability estimator* and a *domain continuation decision maker*, which use information available in the expert itself.

When a new input utterance is received, at Stage 1, the activation probability estimators of all non-activated experts estimate probabilities and send them to the domain selector. Then at Stage 2, the domain selector sends their maximum value to the expert of the preceding domain and asks it to decide whether it continues to deal with the new input utterances or does not continue, or it deals with the utterance as out-of-domain. If it decides not to continue, the domain selector selects the expert that outputs the highest probability at Stage 1.

The reason we use the term "framework" is that it does not specify the details of the algorithm and features used in each domain expert's submodules for domain selection. It rather specifies the interfaces of those submodules. Note that RECSCORE, RECSCORE+BIAS, and MAXPROB can be considered as one of the implementations of this framework. This framework, however, allows developers to use a wider variety of features and gives flexibility in designing those submodules.

5 Example Implementation and Evaluation

Since the proposed framework is an extension of the previous methods, if the activation probability estimator and domain continuation decision maker for each expert are designed well and trained using enough data, it should outperform previous methods that satisfy extensibility. We believe that this theoretical consideration and an experimental result using a human-system dialogue corpus show the viability of the framework. Below we explain our implementation and an experiment.

5.1 Data

For the implementation and evaluation, we used a corpus of dialogues between human users and the World Heritage Site information system described in Section 2.2. Domain selection of this system was performed using hand-crafted rules.

35 participants (17 males and 18 females) whose ages range from 19 to 57 were asked to engage in

domain	preceding domain	training data A	training data B	test data
RU	RU	134	169	145
	QA	51	102	59
	IP	21	16	23
	subtotal	206	287	227
QA	RU	46	55	51
	QA	783	870	888
	IP	59	87	66
	subtotal	888	1,012	1,005
IP	RU	2	1	3
	QA	7	11	18
	IP	311	305	277
	subtotal	320	317	298
OOD	RU	24	19	39
	QA	168	155	183
	IP	66	68	113
	subtotal	258	242	335
total		1,672	1,858	1,865

Table 1: Number of utterances in each domain in the training and test data.

conversation with the system four times. Each session lasted eight minutes. For each utterance, the correct domain or an OOD label was manually annotated. We also annotated its preceding domain, i.e., the domain in which the previous system utterance was made. It can be different from the previous user utterance's domain because of the system's erroneous domain selection. Utterances including requests in two domains at the same time should be given an OOD label but there are no such utterances. We used data from 23 participants (3,530 utterances) for training and those from the remaining 12 participants (1,865 utterances) for testing. We further split the training data into training data A (1,672 utterances) and B (1,858 utterances) to train each of the two submodules. Each training data set includes data from two sessions for each participant. Table 1 shows detailed numbers of utterances in the data sets.

5.2 Implementation

5.2.1 Expert Classes

Among the ten experts, eight IP (Interactive Presentation) experts have the same dialogue strategy and most of the predicted user utterance patterns. In addition, the number of training utterances for each IP expert's domain is small. We therefore used all training utterances in the IP domains to build a common ASR language model (LM), a common activation probability estimator, and a common domain continuation decision maker for all IP experts. Hereafter we call the set of IP experts *the IP expert class*. The RU (Request Understanding) expert and the QA (Question Answer) expert are themselves also expert classes.

expert class	QA	IP	RU
LM for ASR	trigram	trigram	finite-state grammar
language understanding	keyphrase -based	keyphrase -based	finite-state transducer
vocabulary size (word)	1,140	407	79
phone error rate (%)	10.95	19.47	23.60

Table 2: Speech understanding in each expert.

5.2.2 Speech Understanding

For all experts, we used the Julius speech recognizer and the acoustic model in the Japanese model repository (Kawahara et al., 2004).[2] Features of speech understanding in each expert class are shown in Table 2. Compared to the system used for data collection, LMs are enhanced based on the training data. We obtained the ASR performance on the utterances in each domain in the test data in terms of phone error rates. This is because Japanese has no standard word boundaries so it is not easy to correctly compute word error rates. The poor performance of ASR for IP is mainly due to the small amount of training utterances for LM and that for RU is mainly due to out-of-grammar utterances.

5.2.3 Stage 1

For Stage 1, we used logistic regression to estimate the probability that a non-activated expert would be activated by a user utterance. Features for logistic regression include those concerning speech recognition and understanding results as well as dialogue history (see Table 5 for the full list of features). These features are *expert-dependent*. This makes it possible to estimate how the input utterance is suit-

[2] Multiple LMs can be used at the same time with Julius.

able to the dialogue context more precisely than using just features available in any kind of expert.

To train the activation probability estimators, we fitted logistic regression coefficients using Weka data mining toolkit ver.3.6.2 (Witten and Frank, 2005)[3] and training data A. In the training for each expert class, we used utterances whose preceding domain was not that of the class because activation probabilities are estimated only for such utterances during domain selection. If the utterance is in a domain of the expert class, it is assigned an *activate* label and otherwise *not-activate*. Next, we performed feature selection to avoid overfitting. We used backward stepwise selection so that the weighted (by the sizes of *activate* and *not-activate* labels) average of the F_1 scores for training set B could be maximized. Table 6 lists the remaining features and their significances in terms of the F_1 score obtained when each feature is removed. Then, we duplicated the *activate*-labeled utterances in the training data A so that the ratio of *activate*-labeled utterances to *not-activate*-labeled utterances became 1 to 3. This is because the training data include a larger number of *not-activate*-labeled utterances and thus the results would be biased. The ratio was decided by trial and error so that the weighted average of the F_1 scores for training data B becomes high.

5.2.4 Stage 2

For Stage 2, we used multi-class support vector machines (SVMs)[4] to decide if the activated expert should continue to be activated, should not continue, or should regard the input utterance as OOD. We used the same set of features as Stage 1 as well as the maximum activation probability obtained at Stage 1. The training data for the SVM of each expert class is the set of utterances in training data B whose preceding domain is in that expert class, because domain continuation is decided only for such utterances during domain selection. They are labeled *continue*, *not-continue*, or *OOD*. Next, we performed backward stepwise feature selection so that the weighted average of F_1 scores for *continue*, *not-continue*, and *OOD* utterance detection on training data A could be maximized. Remaining features are listed in Table 7. The maximum activation probability was found to be significant in all expert classes. This suggests our two-stage framework that uses maximum activation probability is viable. Then, we duplicated utterances with *not-continue* label and *OOD* label in the training data so that the ratio of *continue*, *not-continue*, and *OOD* utterances became 3:1:1. This is because the number of utterances with the *continue* label is far greater than others. The ratio was experimentally decided by trial and error so that the weighted average of F_1 scores on training data A becomes high.

[3]Multinominal logistic regression model with a ridge estimator with Weka's default values.

[4]Weka's SMO with the linear kernel and its default values.

5.3 Evaluation

5.3.1 Compared Methods

We compared the *full implementation* described in Section 5.2 (FULLIMPL hereafter) with the following four methods which satisfy *extensibility*. Note that the first three methods were mentioned in Section 4.

RECSCORE**:** This chooses the expert class whose recognition score is the maximum (Isobe et al., 2003). We used the ASR acoustic score normalized by the duration of the utterance. If the IP expert class was chosen, the IP expert that had been most recently activated was chosen, because, in this system, domain shifts to other IP experts never occur due to the system constraints and the user did not try to do it. If none of the experts had a higher score than a fixed threshold, it recognized the utterance as OOD. The threshold was experimentally determined using the training data so that the weighted (by the sizes of OOD and non-OOD utterances) average of the F_1 scores of OOD/non-OOD classification is maximized.

RECSCORE+BIAS**:** This is the same as RECSCORE except that a fixed value (bias) is added to the score used in RECSCORE for the expert of the preceding domain. This is basically the same as Lin et al.'s (1999) method but we use a different recognition score since the recognition score they used cannot be used in our system due to the difference of speech understanding methods. The most appropriate bias for each expert class was decided using the training data so that the weighted average of the F_1 scores could be maximized. OOD detection was done in the same way as RECSCORE.

method	class	recall	precision	F_1	weighted ave. F_1
RECSCORE	cont.	0.763	0.867	0.812	
	shift	0.559	0.239	0.335	
	OOD	0.501	0.848	0.630	0.789
RECSCORE +BIAS	cont.	0.917	0.824	0.868	
	shift	0.400	0.421	0.410	
	OOD	0.501	0.848	0.630	0.838
MAXPROB	cont.	0.925	0.843	0.882	
	shift	0.282	0.264	0.273	
	OOD	0.275	0.477	0.348	0.832
NOACTIV PROB	cont.	0.875	0.890	0.882	
	shift	0.464	0.385	0.421	
	OOD	0.785	0.843	0.813	0.849
FULLIMPL	cont.	0.902	0.907	0.904	
	shift	0.591	0.565	0.578	
	OOD	0.824	0.829	0.826	**0.883**
CLASSIFIER (reference)	cont.	0.956	0.881	0.917	
	shift	0.545	0.759	0.635	
	OOD	0.755	0.885	0.815	0.899

Table 3: Evaluation results ("cont." means "continue.").

MAXPROB**:** The activation probabilities for all experts were obtained using logistic regression and the expert whose probability was the maximum was selected. IP experts that had never been activated were excluded because they cannot be activated due to system constraint. For logistic regression, in addition to the features used in FULLIMPL, the previous domain was used as a feature so that domain continuity was taken into account. Feature selection was also performed. The probability that the utterance is OOD was estimated in the same way using the features concerning speech understanding. If the maximum probability of OOD detection was greater than the maximum activation probability, then the utterance was considered to be OOD.

NOACTIVPROB**:** This is the same as FULLIMPL except that Stage 2 does not use the result of Stage 1, i.e., maximum activation probability.

5.3.2 Evaluation Results

To evaluate the domain selection, we focused on domain shifts rather than the selected domain. We classified the domain selection results into domain continuations, domain shifts, and OOD utterance detection. As the evaluation metric, we used the weighted average of F_1 scores for those classes. Here the weight is the ratio of those classes of correct labels. Note that shifting to an incorrect domain is counted as a false positive when calculating precision for domain shifts. Table 3 shows the results. In addition, the confusion matrices for the three best methods are shown in Table 4. We found FULLIMPL outperforms the other four methods. We also found that the differences between the results of the compared methods are all statistically significant ($p < .01$) by two-tailed binomial tests.

For reference, we also evaluated a classifier-based method that uses features from all the experts. Note that this method does not satisfy extensibility because it requires training data in the same set of domains as the target system. We evaluated this just for estimating how well our proposed method works while satisfying extensibility. It classifies each utterance into one of four categories: the QA expert's domain, the RU expert's domain, the most recently activated IP expert's domain, and OOD. If no IP expert has been activated before the utterance, threefold classification was performed. The training and test data were split depending on whether one of the IP experts has been activated before, and training and testing were separately conducted. The training data A was used for training SVM classifiers. Then feature selection was performed using the training data B. The performance of this method is shown as CLASSIFIER in Tables 3 and 4. Although this method outperforms FULLIMPL, FULLIMPL's performance is close to this method. This shows that our method does not degrade its performance very much even though it satisfies extensibility.

5.3.3 Discussion

One of the reasons why FULLIMPL outperforms other methods is that its precision for domain shifts is relatively higher than the other methods. This suggests it can avoid erroneous domain shifts, thus the proposed two-stage framework is more *robust*. RECSCORE+BIAS performed relatively well despite it used only limited features. We guess this is because adding preferences to the preceding domain was effective since domain shifts are rare in these data. Its low F_1 score for OOD utterances suggests using just recognition scores is insufficient to detect them. The comparison of FULLIMPL with NOACTIVPROB shows the effectiveness of using maximum activation probability in the second stage.

The F_1 score for domain shifts is low even with

RECSCORE+BIAS:

correct	estimated result				total
	cont.	correct shift	wrong shift	OOD	
continue	1,201	-	82	27	1,310
shift	115	88	14	3	220
OOD	142	-	25	168	335
total	1,458	88	121	198	1,865

NOACTIVPROB:

correct	estimated result				total
	cont.	correct shift	wrong shift	OOD	
continue	1,146	-	123	41	1,310
shift	92	102	18	8	220
OOD	50	-	22	263	335
total	1,288	102	163	312	1,865

FULLIMPL:

correct	estimated result				total
	cont.	correct shift	wrong shift	OOD	
continue	1,181	-	77	52	1,310
shift	70	130	15	5	220
OOD	51	-	8	276	335
total	1,302	130	100	333	1,865

CLASSIFIER (reference):

correct	estimated result				total
	cont.	correct shift	wrong shift	OOD	
continue	1,252	-	30	28	1310
shift	92	120	3	5	220
OOD	77	-	5	253	335
total	1,421	120	38	286	1,865

Table 4: Confusion matrices for the domain shifts.

FULLIMPL, although it is higher than those with other methods. One typical reason for this is that when one keyword in the ASR result of an utterance to shift the domain is also in the vocabulary of the preceding domain's expert, the selection tends to continue the previous domain by mistake. For example, an utterance "tell me about other World Heritage Sites" to shift from an IP domain to the QA domain is sometimes misclassified as an IP domain utterance, because "World Heritage Sites" is also in IP domains' vocabulary. We think this is because the training data do not include a sufficient amount of utterances that shift domains, and that a larger amount of training data would solve this problem.

6 Concluding Remarks

This paper presented a novel general framework for domain selection in extensible multi-domain spoken dialogue systems. This framework makes it possible to build a robust domain selector because of its flexibility in exploiting features and taking into account domain continuity. An experiment with data collected with an example multi-domain system supported the viability of the proposed framework. We believe that this framework will promote the development of multi-domain spoken dialogue systems and conversational robots/agents.

Among future work is to investigate how accurate the activation probability estimator and the domain continuation decision maker in each domain expert should be for achieving a reasonable accuracy in domain selection. We also plan to conduct experiments with systems that have a larger number of domain experts to verify the scalability of this framework. In addition, we will explore a way to estimate the confidence of the domain selection to reduce erroneous domain selections.

Acknowledgments

The authors would like to thank Hiroshi Tsujino, Yuji Hasegawa, and Hiromi Narimatsu for their support for this research.

References

Hideki Asoh, Toshihiro Matsui, John Fry, Futoshi Asano, and Satoru Hayamizu. 1999. A spoken dialog system for a mobile office robot. In *Proc. 6th Eurospeech*, pages 1139–1142.

Jennifer Chu-Carroll and Bob Carpenter. 1999. Vector-based natural language call routing. *Computational Linguistics*, 25(3):361–388.

Joakim Gustafson and Linda Bell. 2000. Speech technology on trial: Experiences from the August system. *Natural Language Engineering*, 6(3&4):273–286.

Aaron Heidel and Lin-shan Lee. 2007. Robust topic inference for latent semantic language model adaptation. In *Proc. ASRU-07*, pages 177–182.

Bo-June (Paul) Hsu and James Glass. 2006. Style and topic language model adaptation using HMM-LDA. In *Proc. EMNLP '06*, pages 373–381,.

Satoshi Ikeda, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. 2008. Extensibility verification

of robust domain selection against out-of-grammar utterances in multi-domain spoken dialogue system. In *Proc. Interspeech-2008 (ICSLP)*, pages 487–490.

T. Isobe, S. Hayakawa, H. Murao, T. Mizutani, K. Takeda, and F. Itakura. 2003. A study on domain recognition of spoken dialogue systems. In *Proc. Eurospeech-2003*, pages 1889–1892.

Tatsuya Kawahara, Akinobu Lee, Kazuya Takeda, Katsunobu Itou, and Kiyohiro Shikano. 2004. Recent progress of open-source LVCSR engine Julius and Japanese model repository. In *Proc. Interspeech-2004 (ICSLP)*, pages 3069–3072.

Kazunori Komatani, Naoyuki Kanda, Mikio Nakano, Kazuhiro Nakadai, Hiroshi Tsujino, Tetsuya Ogata, and Hiroshi G. Okuno. 2006. Multi-domain spoken dialogue system with extensibility and robustness against speech recognition errors. In *Proc. 7th SIGdial Workshop*, pages 9–17.

Ian R. Lane and Tatsuya Kawahara. 2005. Incorporating dialogue context and topic clustering in out-of-domain detection. In *Proc. ICASSP-2005*, pages 1045–1048.

Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. 2009. Example-based dialog modeling for practical multi-domain dialog system. *Speech Communication*, 51(5):466–484.

Bor-shen Lin, Hsin-ming Wang, and Lin-shan Lee. 1999. A distributed architecture for cooperative spoken dialogue agents with coherent dialogue state and history. In *Proc. ASRU-99*.

Michael F. McTear. 2004. *Spoken Dialogue Technology*. Springer.

Mikio Nakano, Atsushi Hoshino, Johane Takeuchi, Yuji Hasegawa, Toyotaka Torii, Kazuhiro Nakadai, Kazuhiko Kato, and Hiroshi Tsujino. 2006. A robot that can engage in both task-oriented and non-task-oriented dialogues. In *Proc. Humanoids-2006*, pages 404–411.

Mikio Nakano, Kotaro Funakoshi, Yuji Hasegawa, and Hiroshi Tsujino. 2008. A framework for building conversational agents based on a multi-expert model. In *Proc. 9th SIGdial Workshop*, pages 88–91.

Mikio Nakano, Yuji Hasegawa, Kotaro Funakoshi, Johane Takeuchi, Toyotaka Torii, Kazuhiro Nakadai, Naoyuki Kanda, Kazunori Komatani, Hiroshi G. Okuno, and Hiroshi Tsujino. 2011. A multi-expert model for dialogue and behavior control of conversational robots and agents. *Knowledge-Based Systems*, 24(2):248–256.

Hiromi Narimatsu, Mikio Nakano, and Kotaro Funakoshi. 2010. A classifier-based approach to supporting the augmentation of the question-answer database for spoken dialogue systems. In *Proc. 2nd IWSDS*, pages 182–187.

Yoshitaka Nishimura, Shinichiro Minotsu, Hiroshi Dohi, Mitsuru Ishizuka, Mikio Nakano, Kotaro Funakoshi, Johane Takeuchi, Yuji Hasegawa, and Hiroshi Tsujino. 2007. A markup language for describing interactive humanoid robot presentations. In *Proc. IUI'07*, pages 333–336.

Ian O'Neill, Philip Hanna, Xingkun Liu, and Michael McTear. 2004. Cross domain dialogue modelling: an object-based approach. In *Proc. Interspeech-2004 (ICSLP)*, pages 205–208.

Botond Pakucs. 2003. Towards dynamic multi-domain dialogue processing. In *Proc. Eurospeech-2003*, pages 741–744.

Esa-Pekka Salonen, Mikko Hartikainen, Markku Turunen, Jaakko Hakulinen, and J. Adam Funk. 2004. Flexible dialogue management using distributed and dynamic dialogue control. In *Proc. Interspeech-2004 (ICSLP)*, pages 197–200.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.

expert class	Features	
all classes $i = ru$, ip, qa	$F_{i,r1}$	If $\text{SRR}_{i,1}$ is obtained or not
	$F_{i,r2}$	If $\text{SRR}_{i,1}$ contains a filler or not
	$F_{i,r3}$	min (CMs of words in $\text{SRR}_{i,1}$)
	$F_{i,r4}$	avg (CMs of words in $\text{SRR}_{i,1}$)
	$F_{i,r5}$	(acoustic score of $\text{SRR}_{i,1}$) / duration
	$F_{i,r6}$	LM score of $\text{SRR}_{i,1}$
	$F_{i,r7}$	# of words in $\text{SRR}_{i,1}$
	$F_{i,r8}$	# of words in $\text{SRR}_{i,all}$
	$F_{i,r9}$	($F_{i,r5}$ - (acoustic score of $\text{SRR}_{lv,1}$)) / duration
RU	$F_{ru,r10}$	If $\text{SRR}_{ru,1}$ is an affirmative response
	$F_{ru,r11}$	If $\text{SRR}_{ru,1}$ is a denial response
	$F_{ru,r12}$	# of ASR results with LM_{ru}
	$F_{ru,r13}$	If $\text{SRR}_{ru,1}$ contains the name of a World Heritage Site
	$F_{ru,r14}$	max (CMs of words comprising the name of a World Heritage Site)
	$F_{ru,r15}$	ave (CMs of words comprising the name of a World Heritage Site)
	$F_{ru,h1}$	If $\text{SRR}_{ru,1}$ is an affirmative response (Stage 2 only)
	$F_{ru,h2}$	# of turns since this expert is activated
	$F_{ru,h3}$	# of denial responses recognized since this expert is activated
	$F_{ru,h4}$	$F_{ru,h4}/F_{ru,h3}$
	$F_{ru,h5}$	If the previous system utterance is a confirmation request to a user request for starting a presentation
	$F_{ru,h6}$	If the previous system utterance is an utterance to react to a non-understandable user utterance
	$F_{ru,h7}$	If the system has made a confirmation request to a user request for starting a presentation since this expert was activated
	$F_{ru,h8}$	If the system has made an utterance to react to a non-understandable user utterance since this expert was activated
	$F_{ru,h9}$	If the system has made a confirmation request to a user request for starting a presentation before
	$F_{ru,h10}$	If the system has made an utterance to react to a non-understandable user utterance before

expert class	Features	
IP	$F_{ip,r10}$	If the $\text{SRR}_{ip,1}$ is out of database
	$F_{ip,r11}$	$\sum_j$((# of keyphrases in $\text{SRR}_{ip,j}$) / (# of words in $\text{SRR}_{ip,j}$)) / (# of ASR results)
	$F_{ip,r12}$	$\min_i$(# of keyphrase_i in $\text{SRR}_{ip,all}$ / (# of ASR results))
	$F_{ip,r13}$	$\max_i$(# of keyphrase_i in $\text{SRR}_{ip,all}$ / (# of ASR results))
	$F_{ip,r14}$	avg(CM of keyphrase_i in $\text{SRR}_{ip,1}$)
	$F_{ip,r15}$	$\min_i$ (CM of keyphrase_i in $\text{SRR}_{ip,1}$)
	$F_{ip,r16}$	$\max_i$ (CM of keyphrase_i in $\text{SRR}_{ip,1}$)
	$F_{ip,h1}$	If this expert has been activated before
	$F_{ip,h2}$	Same as $F_{ru,h2}$
	$F_{ip,h3}$	If the previous system utterance is the final utterance of the presentation
	$F_{ip,h4}$	If the previous system utterance is an utterance to react to a user interruption
	$F_{ip,h5}$	Same as $F_{ru,h6}$
	$F_{ip,h6}$	If the system has made the final utterance of the presentation since this expert was activated
	$F_{ip,h7}$	If the system has made an utterance to react to a user interruption since this expert was activated
	$F_{ip,h8}$	Same as $F_{ru,h8}$
	$F_{ip,h9}$	If the system has made the final utterance of the presentation before
	$F_{ip,h10}$	If the system has made an utterance to react to a user interruption before
	$F_{ip,h11}$	Same as $F_{ru,h10}$
QA	$F_{qa,r10}$	Same as $F_{ip,r12}$
	$F_{qa,r11}$	Same as $F_{ip,r13}$
	$F_{qa,r12}$	Same as $F_{ip,r14}$
	$F_{qa,r13}$	Same as $F_{ip,r15}$
	$F_{qa,r14}$	Same as $F_{ip,r16}$
	$F_{qa,r15}$	Same as $F_{ip,r17}$
	$F_{qa,r16}$	If $\text{SRR}_{qa,1}$ is an acknowledgment
	$F_{qa,h1}$	Same as $F_{ru,h1}$
	$F_{qa,h2}$	Same as $F_{ru,h2}$
	$F_{qa,h3}$	Same as $F_{ru,h3}$
	$F_{qa,h4}$	$F_{qa,h4}/F_{qa,h3}$
	$F_{qa,h5}$	If the previous system utterance is the final utterance of an answer
	$F_{qa,h6}$	Same as $F_{ru,h6}$
	$F_{qa,h7}$	If the system has made the final utterance of an answer since this expert was activated
	$F_{qa,h8}$	Same as $F_{ru,h8}$
	$F_{qa,h9}$	If the system has made the final utterance of an answer before
	$F_{qa,h10}$	Same as $F_{ru,h10}$

$\text{SRR}_{i,j}$ means j-th speech recognition result with the language model (LM) for expert class i. $\text{SRR}_{i,all}$ means all the recognition results in the n-best list. F_{i,r_x} are speech understanding related features and F_{i,h_x} are dialogue history related features. $\text{SRR}_{lv,j}$ is an ASR result with a large-vocabulary (60,250 words) statistical model (Kawahara et al., 2004), which we used for utterance verification. CM means confidence measure.

Table 5: Features used in the experiment.

expert class (F_1 score obtained after feature selection)	remaining features (F_1 score obtained when each feature is removed)
RU(0.948)	$F_{ru,r9}$ (0.922), $F_{ru,h8}$ (0.939), $F_{ru,r5}$ (0.940), $F_{ru,r14}$ (0.941), $F_{ru,r2}$ (0.944), $F_{ru,h9}$ (0.944), $F_{ru,h5}$ (0.944), $F_{ru,r13}$ (0.945), $F_{ru,h10}$ (0.945), $F_{ru,r10}$ (0.946), $F_{ru,r8}$ (0.946), $F_{ru,r7}$ (0.946)
IP(0.837)	$F_{ip,r7}$ (0.771), $F_{ip,r6}$ (0.772), $F_{ip,h9}$ (0.781), $F_{ip,h7}$ (0.781), $F_{ip,h11}$ (0.786), $F_{ip,r4}$ (0.79), $F_{ip,r2}$ (0.799), $F_{ip,r16}$ (0.809), $F_{ip,r5}$ (0.809), $F_{ip,r3}$ (0.809), $F_{ip,h4}$ (0.809), $F_{ip,r9}$ (0.814), $F_{ip,r15}$ (0.833), $F_{ip,r12}$ (0.834), $F_{ip,r13}$ (0.835), $F_{ip,h10}$ (0.836)
QA(0.836)	$F_{qa,r14}$ (0.813), $F_{qa,r7}$ (0.817), $F_{qa,r16}$ (0.817), $F_{qa,r10}$ (0.818), $F_{qa,h6}$ (0.820), $F_{qa,r6}$ (0.822), $F_{qa,r3}$ (0.831), $F_{qa,r5}$ (0.832)

Table 6: Features that remained after feature selection at Stage 1 and their significances in terms of the F_1 score obtained when each feature is removed.

expert class (F_1 score obtained after feature selection)	remaining features (F_1 score obtained when each feature is removed)
RU(0.773)	$F_{ru,r3}$ (0.728), $F_{ru,a}$ (0.737), $F_{ru,h5}$ (0.743), $F_{ru,h1}$ (0.751), $F_{ru,r9}$ (0.754), $F_{ru,h10}$ (0.757), $F_{ru,h8}$ (0.757), $F_{ru,r5}$ (0.758), $F_{ru,r2}$ (0.759), $F_{ru,r13}$ (0.762), $F_{ru,r14}$ (0.763), $F_{ru,h9}$ (0.767), $F_{ru,r15}$ (0.768), $F_{ru,r10}$ (0.768), $F_{ru,h3}$ (0.772)
IP(0.827)	$F_{ip,h5}$ (0.808), $F_{ip,r5}$ (0.809), $F_{ip,r4}$ (0.810), $F_{ip,r6}$ (0.811), $F_{ip,a}$ (0.812), $F_{ip,h4}$ (0.812), $F_{ip,r13}$ (0.813), $F_{ip,h3}$ (0.817), $F_{ip,r15}$ (0.818), $F_{ip,r3}$ (0.818), $F_{ip,h10}$ (0.819), $F_{ip,r12}$ (0.820), $F_{ip,h7}$ (0.821), $F_{ip,r11}$ (0.822), $F_{ip,r10}$ (0.822), $F_{ip,h8}$ (0.822), $F_{ip,h6}$ (0.822), $F_{ip,r2}$ (0.824), $F_{ip,r8}$ (0.824), $F_{ip,h9}$ (0.824), $F_{ip,h2}$ (0.825)
QA(0.873)	$F_{qa,a}$ (0.838), $F_{qa,r5}$ (0.857), $F_{qa,h1}$ (0.859), $F_{qa,r3}$ (0.862), $F_{qa,r6}$ (0.865), $F_{qa,h8}$ (0.867), $F_{qa,r7}$ (0.868), $F_{qa,r15}$ (0.870), $F_{qa,r8}$ (0.870), $F_{qa,h7}$ (0.870), $F_{qa,r12}$ (0.871), $F_{qa,r2}$ (0.871), $F_{qa,r16}$ (0.871), $F_{qa,h4}$ (0.871), $F_{qa,h3}$ (0.871), $F_{qa,r11}$ (0.872), $F_{qa,h6}$ (0.872), $F_{qa,h5}$ (0.872)

Table 7: Features that remained after feature selection at Stage 2 and their significances in terms of the F_1 score obtained when each feature is removed. $F_{ru,a}$, $F_{ip,a}$, and $F_{qa,a}$ are the maximum activation probabilities obtained at Stage 1.