INCORPORATING DIALOGUE CONTEXT AND TOPIC CLUSTERING IN OUT-OF-DOMAIN DETECTION

Ian R. Lane[1,2], *Tatsuya Kawahara*[1,2]

[1]School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
[2]ATR Spoken Language Translation Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

ABSTRACT

The detection and handling of OOD (out-of-domain) user utterances are significant problems for spoken language systems. We have proposed a novel OOD detection framework, which makes use of classification confidence scores of multiple topics. In this paper, we extend this framework in order to handle natural language dialogue. Specifically, two issues are addressed. First, to effectively incorporate dialogue context, we investigate methods to combine multiple utterances at various stages of the OOD detection process. Second, to improve robustness on spontaneous speech, we introduce a topic clustering scheme which provides reliable topic classification confidence even for indistinct utterances. The system is evaluated on natural dialogue via the ATR speech-to-speech translation system, and a significant improvement in OOD detection accuracy was achieved by incorporating the two proposed techniques.

1. INTRODUCTION

Spoken language systems are typically developed specifically to operate over limited and definite domains, as defined by the back-end application system. However, users, especially novice users, do not always have an exact concept of the application domain and may attempt utterances that cannot be handled by the back-end system. These are referred to as OOD (out-of-domain) utterances in this paper.

Most current systems consider all input utterances to be in-domain. This assumption, however, often leads to confusion in users. For example, users can interact via a speech-to-speech translation system as shown in Figure 1. For an in-domain task (Example A), users are able to overcome speech recognition and machine translation errors by re-phrasing the input sentence. However, when users attempt an OOD task (Example B), a deadlock will occur, as translation will fail no matter how the utterance is rephrased. To overcome this problem, OOD utterances must be accurately detected and appropriate feedback should be generated. This will enable users to determine whether to re-attempt the current task after being confirmed as in-domain, or to halt after being informed that it is out-of-domain and cannot be handled by the system.

Research in OOD detection is very limited, and typically focused on single utterance tasks, such as call routing [1, 2]. In [3], we proposed an OOD detection framework based on topic classification and in-domain verification. In this framework, the application domain is assumed to consist of multiple sub-domain topics.

	Example A: In-domain dialogue, re-phrased
JPN	[*Excuse me, I'd like to go to a hotel in town what would be the best way to get there.*]
	Recognition/Translation incorrect
ENG	Pardon me.
JPN	[*Please tell me how to get to a hotel in town.*]
ENG	The easiest way is to take a taxi.
	...

	Example B: Out-of-domain dialogue
ENG	Good Morning, Brown and Associates, how may I help you?
	Recognition/Translation incorrect
JPN	[*Could you say that again?*]
ENG	Yes, this is the office of Brown and Associates.
	Recognition/Translation incorrect
JPN	[*Could you say that again?*]
ENG	Yes, this is Mr. Browns' office, how may I help you?
	...

Fig. 1. OOD dialogue in speech based translation

OOD detection is performed by first calculating confidence scores for each topic class and then applying an in-domain verification model to this vector. The in-domain verification model is trained using deleted interpolation of topics, enabling the system to be developed even when no OOD training data is available. The performance of the system was preliminarily evaluated on a simple travel phrasebook task, where OOD detection was performed on read-speech utterances of prepared sentences.

In this paper, we extend the proposed OOD detection framework to handle natural spoken dialogue. We investigate various methods to incorporate dialogue context into the framework. Compared to the phrasebook task in [3], where sentences are typically related to a single topic, in natural dialogue the relationship between utterances and individual topics is often indistinct. To overcome this problem, we introduce a topic clustering scheme where *meta-topics* are created to provide coverage over closely related topic classes, improving the robustness of topic classification. The effectiveness of these two techniques is evaluated on natural dialogue via a speech-to-speech translation system.

0-7803-8874-7/05/$20.00 ©2005 IEEE

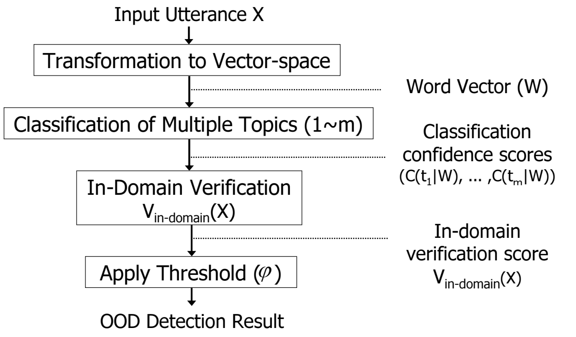

Fig. 2. OOD utterance detection based on topic classification

2. SYSTEM OVERVIEW

In the proposed framework, the training set is initially split into multiple topic classes. In the work described in this paper, topic classes are pre-defined and the training set is hand-labeled appropriately. These data are then used to train the topic classification models.

An overview of the OOD detection framework is shown in Figure 2. Speech recognition is performed by applying a generalized language model that covers all in-domain topics, and a recognition hypothesis X is generated. OOD detection is then performed in the following steps. First, the recognition hypothesis X is transformed to a vector-space representation W and topic classification confidence scores $(C(t_1|W), \dots, C(t_m|W))$ are generated by applying classification models for each topic class t_i. Next, an in-domain verification model $V_{in-domain}(X)$ is applied to the vector of topic classification scores and an in-domain verification score is generated. Finally, an OOD decision is made by applying a threshold φ to this score. We have previously shown in [4] that SVM-based topic classification and linear discriminate verification modeling are suitable for the proposed framework. These are described briefly below.

2.1. SVM-based Topic Classification

Topic classification is based on a vector-space model, where sentences are represented as a vector of occurrence counts, relating to word, word-pair, and word-triplet features. Within this vector-space, SVMs (support vector machines) [5] are trained to discriminate each topic class from others.

Classification is performed by feeding the vector representation W of the input utterance X to each SVM classifier. A classification confidence score ($C(t_i|W)$) is computed by applying a sigmoid function to the resulting SVM distance.

2.2. In-domain Verification

In-domain verification involves applying a linear discriminate model (Equation 1) to the resulting confidence vector from topic classification. The linear discriminant weights $(\lambda_1, \dots, \lambda_m)$ are trained using deleted interpolation of topics as described in [3].

$$V_{in-domain}(X) = \sum_{i=1}^{m} \lambda_i C(t_i|W) \qquad (1)$$

W: vector representation of input utterance X
m: number of topic classes

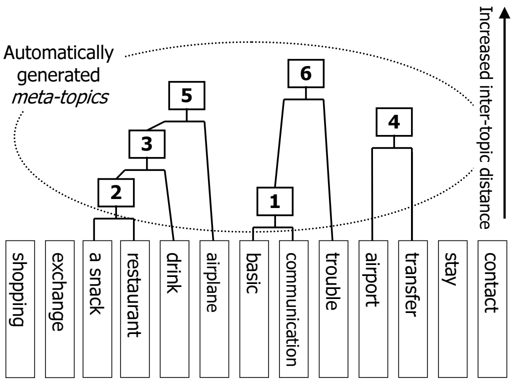

Fig. 3. Topic clustering

3. TOPIC CLUSTERING

In natural dialogue, tasks are often completed through a sequence of utterances. Some utterances may not be full linguistic sentences, and the relationship between utterances and individual topics is often indistinct. To improve topic classification robustness, we introduce a topic clustering scheme, where a set of *meta-topic* classes are generated to provide coverage over closely related and confusable topic classes.

Meta-topics are generated by performing agglomerative clustering to the original in-domain topic classes. Clustering involves iteratively determining the closest topic pairs and merging them until the distances between all topics are greater than some predefined threshold. The distance measure applied during clustering $dist(t_i, t_j)$ is defined as the average distance between topic t_i's training data (S_i) and topic t_j's SVM hyperplane and vice versa (Equation 2).

$$\begin{aligned} dist(t_i, t_j) = &\parallel \underset{W \in S_i}{\text{average}}\, dist_{\perp}(W, t_j) - \underset{W \in S_j}{\text{average}}\, dist_{\perp}(W, t_j) \parallel \\ &+ \parallel \underset{W \in S_j}{\text{average}}\, dist_{\perp}(W, t_i) - \underset{W \in S_i}{\text{average}}\, dist_{\perp}(W, t_i) \parallel \qquad (2) \end{aligned}$$

S_i: set of training sentences of topic class t_i
$dist_{\perp}(W, t_j)$: perpendicular distance from input sentence W to SVM hyperplane of topic t_j

The resulting clustering structure for an evaluation task domain is shown in Figure 3. In this example, six clusters were generated $(1, \dots, 6)$. The lowest layer of the structure corresponds to the individual topic classes and those classes higher in the hierarchy correspond to *meta-topics* that provide coverage over multiple topics. Topic classification models are trained for all individual topics and *meta-topics*, and these models are used to compute the topic confidence vector $C(t_i|W)$ during OOD detection.

4. INCORPORATING DIALOGUE CONTEXT

When applying OOD detection to spoken dialogue, the decision should be made for a sequence of utterances considering dialogue context. Namely, for a set of n consecutive utterances $(X_1, \dots, X_n)$, a single in-domain verification score

$V_{in-domain}(X_{[1,\ldots,n]})$ is calculated. We investigate three methods to incorporate dialogue context into the OOD detection framework, involving combining utterances at three levels: word vector, topic classification, and in-domain verification. These three methods are explained in the following sub-sections.

4.1. Word Vector-level Combination (WRD)

The simplest method is to concatenate the word sequences of multiple utterances $(X_1, \ldots, X_n)$ and generate a single word vector $(W_{[1,\ldots,n]})$ by summing word occurrences over all utterances (Equation 3). Topic classification is then applied to this vector and the resulting scores are used for in-domain verification (Equation 4).

$$W_{[1,\ldots,n]} = \sum_{j=1}^{j \le n} W_j \tag{3}$$

$$V_{in-domain_{avg}}(X_{[1,\ldots,n]}) = \sum_{i=1}^{m} \lambda_i C(t_i | W_{[1,\ldots,n]}) \tag{4}$$

4.2. Topic Classification-level Combination (TOP)

An alternative method is to combine utterances at the topic classification level. Topic classification scores are calculated independently for each utterance $(C(t_i|W_1), \ldots, C(t_i|W_n))$ and then averaged (Equation 5), generating a single topic classification vector. In-domain verification is then applied as shown in Equation 6.

$$C_{avg}(t_i|W_1, \ldots, W_n) = \frac{1}{n} \sum_{j=1}^{j \le n} C(t_i|W_j) \tag{5}$$

$$V_{in-domain_{avg}}(X_{[1,\ldots,n]}) = \sum_{i=1}^{m} \lambda_i C_{avg}(t_i|W_1, \ldots, W_n) \tag{6}$$

4.3. In-domain Verification-level Combination (VER)

In this method, topic classification and in-domain verification is applied independently for each input utterance. The in-domain verification score is then averaged over the individual verification scores (Equation 7).

$$V_{in-domain_{avg}}(X_{[1,\ldots,n]}) = \sum_{j=1}^{j \le n} V_{in-domain}(X_j) \tag{7}$$

5. EXPERIMENTAL EVALUATION

5.1. Experiment Setup

The performance of the proposed OOD detection framework is evaluated for real English/Japanese spoken dialogue via a speech-to-speech translation system, which was developed at ATR [6]. The system consists of statistical machine translation back-ends for English-to-Japanese and Japanese-to-English translation, and user interfaces based on speech recognition and text-to-speech modules. OOD detection systems were integrated into the above system for each language side. The test set consists of 305 dialogue sessions between native English and Japanese speakers for various dialogue scenes.

Table 1. ATR-BTEC training corpus

Domain:	Basic Travel Expressions
Languages:	English, Japanese
Training Set:	14 topics (*accommodation*, *shopping*, ...)
Training Set:	400k sentences
Lexicon Size:	10k/20k (English/Japanese respectively)

Table 2. OOD detection performance for topic clustering

Initiating speaker	OOD Topic	No. Sessions OOD	ID	OOD detection accuracy (EER%) T	C
English	accommodation	37	113	15.6	11.2
	airport	8	144	13.5	13.9
	restaurant	11	142	27.4	25.6
	shopping	11	140	13.0	13.0
	sightseeing	20	131	23.2	15.1
	TOTAL	**87**	**670**	**18.4**	**14.9**
Japanese	accommodation	44	111	27.6	20.6
	airport	9	144	11.1	11.1
	restaurant	8	144	12.5	12.5
	shopping	22	132	23.1	13.6
	sightseeing	20	134	28.4	24.8
	TOTAL	**103**	**665**	**22.1**	**17.3**

T: classifiers applied for original topics only
C: classifiers for topic clustered *meta-topics* included

The performance of the OOD detection framework was evaluated for 5 test scenarios. For each scenario, one topic was set as OOD of the system, and the language model for speech recognition and OOD detection modules were trained with the remaining in-domain topic data from the ATR-BTEC corpus (Table 1) [7].

System performance was evaluated using the EER (equal error rate) measure. The OOD detection threshold (φ) was selected such that the FAR (false acceptance rate) and FRR (false rejection rate) were equal. FAR is the percentage of falsely accepted OOD sessions, and FRR is the percentage of falsely rejected in-domain dialogue sessions.

5.2. Evaluation of Topic Clustering

First, we evaluate the effectiveness of the proposed topic clustering scheme. In this experiment, OOD detection was applied to the correct transcriptions of the initial $(n = 1)$ utterance of each dialogue. The performance for the English and Japanese dialogue sides is shown for the five test scenarios in Table 2. For each test scenario, one topic was set as OOD of the system (Table 2, column 2) and the remaining topics were considered as in-domain. The OOD detection accuracy when only the original topic classifiers were applied (*T*) and when clustering was conducted (*C*) are shown.

Topic clustering provides a total reduction in EER of 3.5 points (from 18.4% to 14.9%) and 4.8 points (from 22.1% to 17.3%) for the English and Japanese sides, respectively. We observed that even when an exact topic could not be identified for in-domain utterances, confidence scores of the *meta-topic* classes provided evidence that the utterance was in-domain.

Table 3. Evaluation of utterance combination

Initiating speaker	Combination method	OOD detection accuracy (EER%)		
		$n = 1$	$n = 2$	$n = 3$
English	WRD	18.4	22.9	17.7
	TOP	-	18.8	16.5
	VER	-	21.7	21.1
Japanese	WRD	22.1	21.8	21.6
	TOP	-	20.8	20.2
	VER	-	24.4	24.7

WRD:	word vector-level combination
TOP:	topic classification-level combination
VER:	in-domain verification-level combination

Table 4. Speech recognition accuracy for test data

Language	In-domain		Out-of-domain	
	WER	SER	WER	SER
Japanese	21.2%	47.0%	23.8%	54.2%

5.3. Evaluation of Utterance Combination

Next, we investigate the system performance when dialogue context is incorporated. We compare three methods to combine multiple utterances as described in Section 4. The system performance when applied to correct transcriptions is shown in Table 3. The performance of each method was evaluated for various numbers of utterances, $n = (1, 2, 3)$.

Combining utterances at the topic classification-level provided the best performance with a reduction in EER of 1.9 points, for both the English and Japanese sides ($n = 3$). However, this improvement is relatively small, suggesting that OOD detection tasks tend to be dominated by the initial utterance.

Utterance combination at the word vector and in-domain verification level degraded the detection accuracy. At the word vector-level, a shift in topic within a single session cannot be correctly handled by a single vector, thus combining utterances at this level is unsuccessful. At the in-domain verification-level, the dynamic range of the verification scores is large, so averaging the scores over multiple utterances tends to be affected by outliers.

5.4. Overall System Performance

Finally, topic clustering and utterance combination (at the topic classification-level) were combined and the system was evaluated when applied to both the correct transcriptions and ASR results. The average WER for the Japanese dialogue side for the in-domain and OOD sets is shown in Table 3. As the English ASR is still under development, we did not integrate it in this work.

The OOD detection performance for the original OOD framework, and when topic clustering and dialogue context are incorporated are shown in Figure 5. A significant reduction in detection errors is gained for the transcription case. The clustering and utterance combination techniques provided a reduction in EER of 4.8 and 1.9 points individually, and when combined a total reduction in EER of 6.4 points (from 22.1% to 15.7%) was gained for the $n = 3$ case. Some degradation for the ASR case is observed (especially for the $n = 2$ and $n = 3$ cases). However considering the WER of 20%, the proposed OOD detection approach is robust against speech recognition errors.

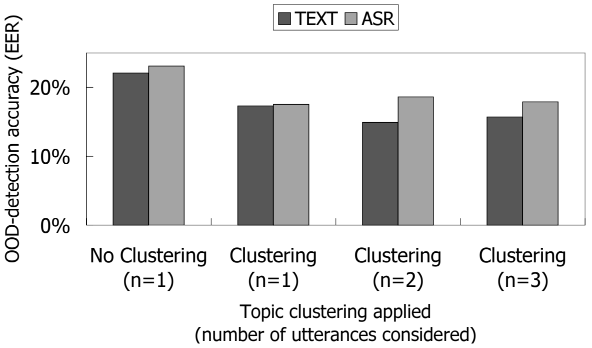

Fig. 4. Combined performance on transcriptions and ASR results

6. CONCLUSIONS

We have investigated OOD detection for natural spoken dialogue by incorporating dialogue context. To improve system robustness, we also introduced topic clustering. The performance of the proposed techniques was evaluated on real dialogue via a speech-to-speech translation system. Topic clustering significantly improved OOD detection performance and a small improvement was also gained by combining multiple utterances during topic classification. The system performance on ASR results was similar to that for transcriptions, showing that the proposed framework works robustly against speech recognition errors.

Acknowledgements: The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled, ”A study of speech dialogue translation technology based on a large corpus”.

7. REFERENCES

[1] A. Gorin, G. Riccardi, and J. Wright. Automated natural spoken dialogue. In *IEEE Computer Magazine, vol. 35, no. 4, pp. 51-56*, April 2002.

[2] P. Haffner, G. Tur, and J. Wright. Optimizing SVMs for complex call classification. In *ICASSP*, 2003.

[3] I. Lane, T. Kawahara, T. Matsui, and S. Nakamura. Out-of-domain detection based on confidence measures from multiple topic classification. In *Proc. IEEE-ICASSP*, 2004.

[4] I. Lane, T. Kawahara, T. Matsui, and S. Nakamura. Topic classification and verification modeling for out-of-domain utterance detection. In *Proc. ICSLP*, 2004.

[5] T. Joachims. Text categorization with support vector machines. In *Proc. European Conference on Machine Learning*, 1998.

[6] T. Takezawa, A. Nishino, K. Takashima, T. Matsui, and G. Kikui. An experimental system for collecting machine-translation aided dialogues. In *Proc. FIT2003, Vol. 2, pp. 161-162*, 2003.

[7] T. Takezawa, M. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Towards a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. LREC, pp. 147-152*, 2002.