社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－SLP－54（50）
2004／12／22

対話コンテキストとトピッククラスタリングを用いたドメイン外発話の検出

レーン　イアン[†,††]　河原　達也[†,††]　中村　哲[††]

† 京都大学 情報学研究科 知能情報学専攻
〒606–8501 京都市左京区吉田二本松町
†† ATR 音声言語コミュニケーション研究所
〒619–0288 京都府相楽郡精華町光台 2-2-2
E-mail: †ian@ar.media.kyoto-u.ac.jp

あらまし　音声言語システムにおいて、システムが想定しない発話（ドメイン外発話）の検出が重要な問題である。我々は以前、トピック分類の信頼度とドメイン内検証モデルを用いた検出の枠組みを提案した。本稿では、自然対話を扱えるように２つの手法を導入する。まず、対話コンテキストを導入するため、文、トピック分類、ドメイン内検証の信頼度の３段階のレベルで複数の発話を結合する手法を検討した。さらに、話し言葉音声に対するシステムの頑健性を向上させるためにトピッククラスタリングを導入する。この手法を用いることでトピックが明確ではない発話でも有用なトピック信頼度を得ることができる。ATR の音声翻訳システムを介した自然対話によりシステムの評価を行い、二つの手法を用いることで、ドメイン外発話の検出精度を改善することができた。

キーワード　音声認識, ドメイン外発話の検出, 信頼度尺度, 対話コンテキスト、トピッククラスタリング

Out-of-Domain Detection Incorporating Dialogue Context and Topic Clustering

Ian R. LANE[†,††], Tatsuya KAWAHARA[†,††], and Satoshi NAKAMURA[††]

† Graduate School of Informatics, Kyoto University,
Yoshida-Nihonmatsu-cho, Sakyo-ku, Kyoto 606–8501, Japan
†† ATR Spoken Language Translation Laboratories,
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619–0288, Japan
E-mail: †ian@ar.media.kyoto-u.ac.jp

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: at the sentence, topic classification, and in-domain verification level. 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 was 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.

Key words　Speech recognition, out-of-domain detection, confidence measures, dialogue context, topic clustering

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 on OOD detection is limited, and conventional studies have typically focused on using recognition confidences for rejecting erroneous recognition outputs (e.g., [1], [2]). In these approaches there is no discrimination between in-domain utterances that have been incorrectly recognized and OOD utterances, and thus effective user feedback cannot be generated.

One area where OOD detection has been conducted is call routing tasks [3]~[5]. In these approaches classification models are trained for each call destination, and a garbage model is explicitly trained to detect OOD utterances. To effectively train these models, a large amount of real-world data is required, consisting of both in-domain and OOD training examples. However, reliance on OOD training data is problematic: first, an operational on-line system is typically required to gather such data, and second, it is difficult to gain an appropriate distribution of data that will provide sufficient coverage over all possible OOD utterances.

In previous work [6], we proposed an OOD detection framework based on topic classification and in-domain verification and introduced a training scheme based on deleted interpolation of topics. This training scheme enables the system to be developed when explict OOD training data is not available. In the proposed framework, the application domain is assumed to consist of multiple sub-domain topics. OOD detection is performed by first calculating confidence scores for each topic class and then applying an in-domain verification model to this vector. In [6], the performance of this framework was evaluated on a simple phrasebook task, where OOD detection was performed on individual, read-speech utterances of prepared sentences.

	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.*]
	Translation successful
ENG	The easiest way is to take a taxi.
JPN	[*Where is the taxi stand?*]
	Translation successful
ENG	Go out exit "C" and you'll see it right in front.
	...

	Example B: Out-of-domain dialogue
ENG	Good Morning, Brown and Associates, how may I help you?
	Recognition/Translation incorrect
JPN	[*Could you please 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 via speech based translation

In this paper, we extend the proposed OOD detection framework to handle natural spoken dialogue. Compared to call routing or the phrasebook task, where sentences are typically related to a single topic, in natural dialogue tasks are often completed over multiple utterances, and thus the relationship between utterances and individual topics is often indistinct. For example, individual utterances may not be full sentences or multiple topics may be present in a single utterance. To overcome these problems we investigate two approaches. First, we compare various methods to incorporate dialogue context into the framework. Secondly, to improve the robustness of topic classification, we introduce a topic clustering scheme where *meta-topics* are created to provide coverage over closely related topic classes. The effectiveness of these two techniques is evaluated on natural dialogue via a speech-to-speech translation system.

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 predefined and the training set is hand-

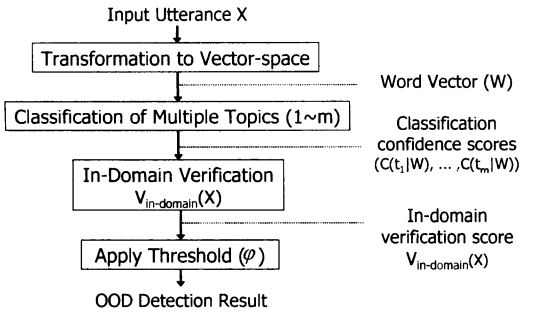

Fig. 2 OOD utterance detection based on topic classification

labeled appropriately. This data is then used to train the topic classification models. Topic classification can also be incorporated during decoding by applying topic-dependent language modeling. We demonstrated the effectiveness of such an approach in [7].

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 $(C(t_1|W), \ldots, C(t_m|W))$ 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 [8] 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) [9] 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 $(C(t_i|W))$ in the range $[0, 1]$ is then 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 vector of topic classification confidence scores generated during topic classification. The linear discriminant weights $(\lambda_1, \ldots, \lambda_m)$ are trained using a deleted interpolation of topics training scheme as described section 2.3.

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

W: vector representation of input utterance X

m: number of topic classes

Table 1 Deleted Interpolation based Training

for each topic i in $[1, M]$
    set topic i as temporary OOD
    set remaining topic classes as in-domain
    calculate $(\lambda_1, \ldots, \lambda_M)$ using GPD (λ_i excluded)
average $(\lambda_1, \ldots, \lambda_M)$ over all iteration

2.3 Verification Model Training using Deleted Interpolation of Topics

In [6], we introduced a method to train the linear discriminant weights $(\lambda_1, \ldots, \lambda_m)$ of the in-domain verification model using only in-domain data. The proposed method combines the GPD (gradient probabilistic descent) algorithm [10] and deleted interpolation. An overview of the training approach is given in Table 1. During training, each topic is iteratively set to be temporarily OOD, and the classifier corresponding to this topic is removed from the verification model. The discriminant weights of the remaining topic classifiers are then estimated using GPD. In this step, the temporary OOD data are used as negative training examples, and a balanced set of the remaining topic classes are used as positive (in-domain) examples. Upon completion, the final model weights are calculated by averaging over all interpolation steps.

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 pre-defined 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

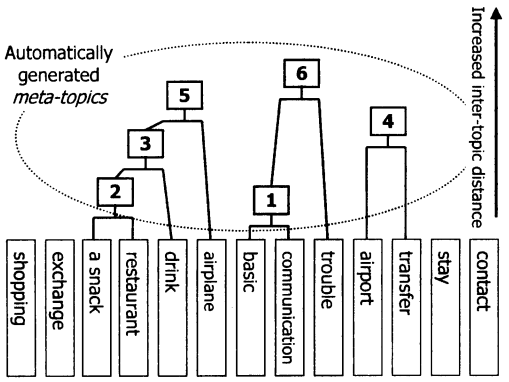

Fig. 3 Topic clustering

topic t_j's SVM hyperplane and vice versa (Equation 2).

$$dist(t_i, t_j) = \| \underset{W \in S_i}{\text{average}}\, dist_\perp(W, t_j) - \underset{W \in S_j}{\text{average}}\, dist_\perp(W, t_j) \| + \| \underset{W \in S_j}{\text{average}}\, dist_\perp(W, t_i) - \underset{W \in S_i}{\text{average}}\, dist_\perp(W, t_i) \| \quad (2)$$

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 $(1, \ldots, 6)$. 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.

As the number of topic classification models is increased due to the introduction of the *meta-topics*, the verification model must be updated to match this. In this case the verifier consists of a set of $m + n$ linear discriminate weights $(\lambda_1, \ldots, \lambda_m, \lambda_{m+1}, \ldots, \lambda_{m+n})$, where n is the number of *meta-topics* introduced. The weights $(\lambda_1, \ldots, \lambda_m)$ correspond to the single topic classifiers, and weights $(\lambda_{m+1}, \ldots, \lambda_{m+n})$ are applied to the *meta-topic* classifiers. These weights are trained using deleted interpolation of topics as described in section 2.3. However, during training, classifiers that relate *meta-topic* which are parents of current temporary OOD topic (t_i) must also be removed during training.

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 $(X_1, \ldots, X_n)$, 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).

$$W_{[1,\ldots,n]} = \sum_{j=1}^{j \leqq n} W_j \quad (3)$$

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

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 (Equation 6).

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

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

4.3 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).

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

Table 2 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)

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 [11]. 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.

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) [12].

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 OOD detection performance for topic clustering

Initiating speaker	OOD Topic	No. Sessions OOD	No. Sessions ID	OOD detection accuracy (EER%) T	OOD detection accuracy (EER%) 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

Table 4 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

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 com-

Table 5 Speech recognition accuracy for test data

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

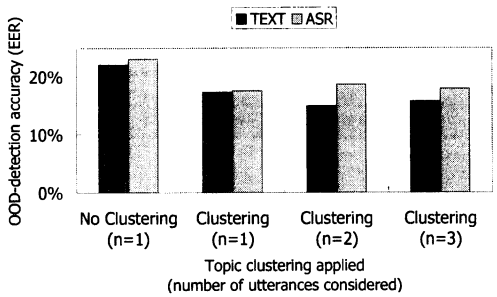

Fig. 4 Combined performance on transcriptions and ASR results

bining 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.

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".

References

[1] T. Hanzen, S. Seneff, and J. Polifroni. Recognition confidence and its use in speech understanding systems. In *Computer Speech and Language*, 2002.

[2] S. J. Cox and S Dasmahapatra. High-level approaches to confidence estimation in speech recognition. In *IEEE Transactions on Speech and Audio*, Vol. 10, No. 7, pp. 460-471, 2001.

[3] C Ma, M. Randolph, and J. Drish. A support vector machines-based rejection technique for speech recognition. In *ICASSP*, 2001.

[4] J. Kuo and C-H Lee. Discriminative training of natural langauge call routers. In *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 1, pp. 24-35, Jan., 2003.

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

[6] 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.

[7] I. Lane, T. Kawahara, and T. Matsui. Language model switching based on topic detection for dialog speech recognition. In *ICASSP*, 2003.

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

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

[10] S. Katagiri, C.-H. Lee, and B.-H. Juang. Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method. In *Proc. IEEE*, vol. 86, pp. 2345-2373, Nov., 1998.

[11] 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.

[12] 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.