# Out-of-Domain Utterance Detection Using Classification Confidences of Multiple Topics

Ian Lane, *Member, IEEE*, Tatsuya Kawahara, *Member, IEEE*, Tomoko Matsui, *Member, IEEE*, and Satoshi Nakamura, *Member, IEEE*

*Abstract*—One significant problem for spoken language systems is how to cope with users' out-of-domain (OOD) utterances which cannot be handled by the back-end application system. In this paper, we propose a novel OOD detection framework, which makes use of the classification confidence scores of multiple topics and applies a linear discriminant model to perform in-domain verification. The verification model is trained using a combination of deleted interpolation of the in-domain data and minimum-classification-error training, and does not require actual OOD data during the training process, thus realizing high portability. When applied to the "*phrasebook*" system, a single utterance read-style speech task, the proposed approach achieves an absolute reduction in OOD detection errors of up to 8.1 points (40% relative) compared to a baseline method based on the maximum topic classification score. Furthermore, the proposed approach realizes comparable performance to an equivalent system trained on both in-domain and OOD data, while requiring no OOD data during training. We also apply this framework to the "*machine-aided-dialogue*" corpus, a spontaneous dialogue speech task, and extend the framework in two manners. First, we introduce topic clustering which enables reliable topic confidence scores to be generated even for indistinct utterances, and second, we implement methods to effectively incorporate dialogue context. Integration of these two methods into the proposed framework significantly improves OOD detection performance, achieving a further reduction in equal error rate (EER) of 7.9 points.

*Index Terms*—Confidence measures, out-of-domain (OOD) utterance detection, speech recognition, spoken language understanding, topic classification, topic clustering.

## I. INTRODUCTION

INTERACTIVE spoken language systems provide a natural and effective interface to a wide range of services, and in recent years have become more prevalent within society. Examples of systems include: spoken dialogue systems for

information access [1]–[4], speech-based call-routing systems [5]–[7] and, more recently, speech-to-speech translation systems [8]–[10]. These systems typically consist of a speech recognition front-end and a natural language processing or information access back-end, for example, a database query module, in spoken dialogue systems.

To operate effectively and realize robust speech recognition, systems are specifically designed to operate over a limited and definite domain, as defined by the back-end application. For users, however, the exact definition of the application domain is not necessarily clear, and users, especially novice users, often attempt utterances that cannot be handled by the backend system. These are referred to as out-of-domain (OOD) utterances in this paper. The definition of OOD is dependent on the type of spoken language system. Definitions for three typical systems are given in Table I. Utterances that can be handled by the back-end system, including discourse utterances, such as "yes," "no," "good morning," or "good bye," are classified as in-domain.

For an effective user interface, systems must provide feedback to the user, informing them when an OOD utterance is encountered. This will enable users to determine whether to continue the current task after being confirmed as in-domain, or to halt, after being informed that it is OOD and cannot be handled by the back-end system. In order to identify OOD utterances, systems must both predict and detect such utterances. In order to predict OOD utterances, the language model must provide some margin in its coverage, such as applying statistical language models, rather than rigid grammar-based models, and a mechanism is required to detect OOD utterances. It is this latter aspect that we address in this paper.

In [10] a speech-to-speech translation "*phrasebook*" system is described that can translate phrases that users are likely to encounter or use while traveling overseas. By incorporating out-of-domain detection into this system, it is able to interact with users as shown in Fig. 1. The first example [Fig. 1(a)] relates to the travel domain, but could not be accurately translated by the back-end system; in such cases, as the utterance is in-domain, the user is requested to re-phrase the utterance. In the case of an OOD utterance [Fig. 1(b)], however, no-matter how the input utterance is re-phrased, it cannot be successfully translated. To handle such utterances, the system must detect that the utterance is OOD, inform the user that the current task cannot be handled by the system, and provide a list of tractable topics, enabling users to switch to an in-domain task.

Conventional research on OOD detection is limited. Previous studies have typically focused on rejecting erroneous recognition outputs based on recognition confidence [11]–[13],

TABLE I
DEFINITION OF OUT-OF-DOMAIN (OOD) FOR VARIOUS SYSTEMS

| System | Out-of-Domain definition |
|---|---|
| Spoken Dialogue | User's query does not relate to back-end information source |
| Call Routing | User's query does not relate to any call destination |
| Speech-to-Speech Translation | Translation system does not provide coverage for referred topic |

| **Translation unsuccessful → re-phrase** | |
|---|---|
| USER | *"Excuse me, I'd like to go to a hotel in town what would be the best way to get there."* |
| **In-domain: Translation unsuccessful** | |
| SYS | Please re-phrase that |
| USER | *"Please tell me how to get to a hotel in town."* |
| **Translation successful** | |
| | ... |

(a)

| **Task OOD → provide feedback to user** | |
|---|---|
| USER | *"How do I print this Word file double-sided?"* |
| **Out-of-Domain** | |
| SYS | I'm sorry, only travel related topics can be handled |
| | ... |

(b)

Fig. 1. Speech based translation for in-domain and OOD tasks. (a) Dialogue example for an in-domain task. (b) Dialogue example for an out-of-domain task.

or confidence measures at the parsing or concept levels [14], [15]. These approaches are based on the assumption that all input utterances will be in-domain and typically provide simple prompts such as "*Please say that again*" or "*I did not understand you, please re-phrase that.*" As there is no discrimination between in-domain utterances that have been erroneously recognized and OOD utterances, these approaches cannot generate effective user feedback, and users cannot determine why the system has failed.

One area where OOD detection has been successfully applied is in automatic call-routing systems. Handling OOD utterances effectively is a requirement for these systems as a large number of calls tend to be OOD. For example in the AT&T "How may I help you" system [5] and the "OASIS" call-steering system [6], around 20% of calls were out-of-domain. These systems typically consist of a speech recognition front-end and a call-classification back-end. Classification methods applied include: Latent Semantic Analysis (LSA) [16], [17], Support-Vector-Machines (SVM) [18]–[20], and other approaches, such as, boosting [21], [22], Naive Bayesian [23], [24], and cosine distance based methods [25]. Utterances are classified as out-of-domain if they are not related to any of the pre-defined call destinations. These calls are forwarded to a human operator. Methods to detect such utterances include confidence-based approaches, where the confidence of the best two classes are compared [25], and approaches that explicitly model OOD utterances [6], [19]. Explicitly modeling OOD utterances is a more effective approach, however, the collection of task-specific OOD training data is problematic; first, a fully operational system is typically required to gather relevant OOD data; second, collecting an adequate distribution of data which

provides coverage over all possible OOD tasks is difficult; and third, OOD data tends to be environment and task-dependent, thus sharing training data between tasks is not possible.

To overcome these problems, we propose a novel OOD detection approach based on topic classification and in-domain verification that can be developed without requiring OOD training data. In the proposed approach, the application domain of the system is assumed to consist of multiple sub-domain topic-classes. OOD detection is performed by first calculating classification confidence scores for all of these classes, and then applying an in-domain verification model to the resulting confidence vector. This generates a final binary in-domain / OOD decision. We also propose a novel method to train the verification model based on deleted interpolation and minimum-classification-error training. This enables the system to be developed using only in-domain data.

To apply the proposed framework to spontaneous spoken dialogue, methods are required to improve system robustness. First, to improve the robustness of topic classification we investigate a topic clustering scheme which enables the system to back-off to a cluster of topics when the exact individual topic is unclear; second, we investigate methods to incorporate dialogue context into the OOD detection process. We evaluate the proposed framework on two tasks: a "*phrasebook*" system, which performs Japanese-to-English translation on individual read-style spoken utterances and the ATR "*machine aided dialogue*" system, which enables speakers of different languages to interact using spontaneous dialogue via parallel speech-to-speech translation systems.

The remainder of this paper proceeds as follows. In Section II we present an overview of the proposed out-of-domain detection framework. The two main elements of this framework, topic classification and in-domain verification, are described in Sections III and IV, respectively. In Section IV-A, we describe in detail the verification model training scheme based on deleted interpolation of topics. In Section V, we investigate methods to extend the framework to handle natural spoken dialogue and focus on methods to incorporate dialogue context into the proposed framework. In Section VI, we present experimental results for the two tasks described above. Conclusions are presented in Section VII.

## II. OVERVIEW OF PROPOSED OOD DETECTION FRAMEWORK

In the proposed framework, OOD detection is performed in two stages: a topic classification stage where confidence scores are generated for a set of in-domain topic classes, and a verification stage that makes the final in-domain/OOD decision. To apply this framework to a particular system, we must first define a set of sub-domain topic-classes, for example, call destinations in call-routing, sub-topics in translation systems, or sub-domains in dialogue systems. In the work described in this paper, we predefined a set of topic classes explicitly and hand-labeled the training set appropriately. This data is then used to train the topic classification models. We have previously demonstrated in [26] that topic classification can also be applied during speech recognition to improve automatic speech recognition (ASR) performance by applying topic-dependent language modeling.
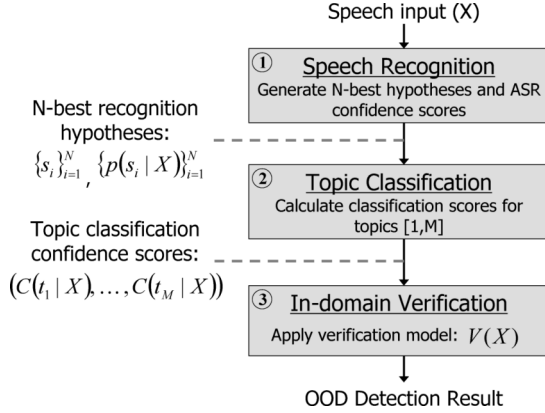
Fig. 2.　OOD utterance detection based on topic classification confidence.



Fig. 3.　Topic classification example.

An overview of the proposed OOD detection framework is shown in Fig. 2. First, speech recognition is performed, applying a language model that provides coverage over all in-domain topic classes, and N-best recognition hypotheses $\{s_1, \ldots s_N\}$ are generated. Next, topic classification confidence scores $(C(t_1|X), \ldots, C(t_M|X))$ are generated based on these hypotheses. Finally, a binary in-domain/OOD decision is generated by applying an in-domain verification model $V(X)$ to this vector. The performance of the proposed approach is dependent on both the accuracy of topic classification and the discriminative ability of the in-domain verification model. These two elements are described in detail in the following sections.

## III. TOPIC CLASSIFICATION

Topic classification (Fig. 2, stage 2), involves calculating a topic-classification confidence vector $(C(t_1|X), \ldots, C(t_M|X))$ for an input utterance X. Each component of this vector consists of a confidence score for a specific topic class $t_j$, which is calculated based on the N-best list of speech recognition hypotheses.

An overview of the topic classification procedure is shown in Fig. 3. First, classification features are extracted from the input sentence $s$ and a feature-vector $(W)$ is generated. Each component of this vector relates to the occurrence of a specific feature: a word, word pair or word triplet. Next, a set of SVM (support vector machine) classification models (one for each topic class $t_1, \ldots, t_M$) are applied to $W$ and confidence scores $C(t_j|W)$ in the range [0,1] are calculated. When applying topic classification to a speech recognition (ASR) result, the confidence vector $(C(t_1|X), \ldots, C(t_M|X))$ is calculated as a weighted linear combination of the individual confidence scores for each N-best hypothesis $\{s_1, \ldots, s_N\}$.

In the following subsections, we describe in detail the topic classification procedure. We describe the extraction of classification features in Section III-A, SVM-based topic classification in Section III-B, and the approach used to apply topic classification to N-best recognition hypotheses in Section III-C. To improve the robustness of topic classification when applied to spontaneous dialogue speech we introduce a topic clustering scheme which generates a set of *meta-topic* classes that provide coverage over multiple topics. During topic classification,
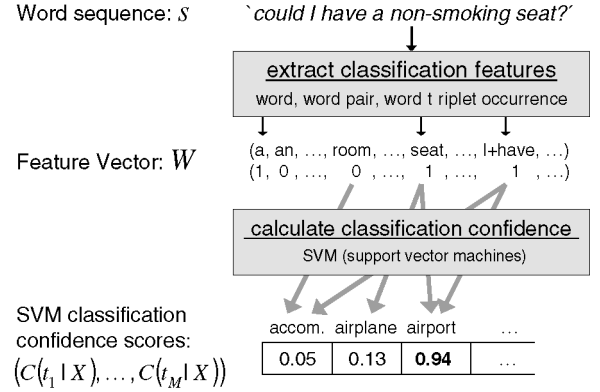
these *meta-topic*s enable the system to determine the cluster an utterance belongs to even when the exact individual topic cannot be identified. We describe this topic clustering scheme in Section III-D.

### A. Feature Extraction

Feature extraction involves generating a vector of feature occurrence counts, $W = (w_1, w_2, \ldots, w_K)$, for an input sentence $s$. First, word tokens, either word baseform (word token with no tense information; all variants are merged), full-word (surface form of words including variants), or word+POS (part-of-speech) information are extracted from the input sentence. Next, a feature vector, $W$, is generated by counting the occurrences of individual word tokens, or $n$-gram features, where an $n$-gram is an agglomeration of $n$ successive tokens. Appropriate cutoffs are applied during training to remove features with low occurrence. In the experimental evaluation (Section VI), we investigate the performance of the OOD detection framework for various feature sets.

Applying a stop-list of very common words during feature extraction can improve topic classification accuracy, however, such lists are highly task dependent, and must be hand tuned for effective performance [5], [25]. To enable task independence, we do not apply a hand-tuned stop-list, but via SVM-based training, those features with little discriminative ability are ignored or given negligibly small weights in the classification model.

### B. SVM-Based Topic Classification

The performance of the proposed OOD detection framework is dependent on the accuracy and robustness of the topic classification method. A large number of classification schemes have successfully been applied to topic classification. Popular methods include: Naive Bayesian classifiers [23], [24], Latent Semantic Analysis (LSA) [16], [17], and Support Vector Machines (SVM) [18]–[20]. In this paper, we explore the use of SVM. SVM is a popular classification technique based on margin maximization. We adopt SVM for the following reasons. First, SVM is appropriate for classification tasks which consist of sparse high-dimensional feature vectors, such as topic classification based on word occurrence features. Second,

SVM performs classification based on a large number of relevant features rather than relying on a limited set of keywords, which improves robustness even when specific keywords are erroneously recognized. Finally, margin maximization based training inherently incorporates robustness against speech recognition errors.

In the proposed framework, topic classification models are trained independently for each topic class. Based on a specific feature set, a discriminative SVM hyperplane $H_j$ is trained for each topic class $t_j$ using a one-vs-others scheme [19], where sentences labeled with the current topic $(t_j)$ are used as positive examples and the remainder of the training set is used as negative training examples. Since this space is very high dimensional (up to 70 000 features, when word-pair and word-triplet features are included) a linear kernel is adequate for classification.

Topic classification is performed by comparing the feature-vector of the input sentence $W$, to each SVM hyperplane. A score for topic $t_j$, $dist_\perp(W, H_j)$ is calculated as the perpendicular distance between $W$ and topic $t_j$'s hyperplane $(H_j)$. This value is positive if $W$ is in-class, and negative otherwise. A confidence score $c_{SVM}(t_j|W)$ is then calculated by applying a sigmoid function $(sigmoid[])$ to this distance

$$c_{SVM}(t_j|W) = \text{sigmoid}\left[\text{dist}_\perp(W, H_j)\right] \qquad (1)$$

where

$$\text{sigmoid}[x] = \frac{1}{1 + e^{(\alpha x + \beta)}} \qquad (2)$$

In the experimental evaluation in Section VI, values of $\alpha = -1.0$ and $\beta = 0.0$ were applied based on preliminary experiments.

### C. Topic Classification for Spoken Utterances

When applying topic classification to an ASR result, the confidence vector $(C(t_1|X), \ldots, C(t_M|X))$ is calculated as a weighted linear combination of the confidence scores for each N-best hypothesis $\{s_1, \ldots, s_N\}$. First, topic classification is applied independently to the feature-vector representation of each hypothesis, generating a topic confidence vector for that hypothesis $(c_{SVM}(t_1|W_i), \ldots, c_{SVM}(t_M|W_i))$. These scores are then linearly combined by weighting each with the posterior probability of that recognition hypothesis. Using this approach the confidence score of topic class $t_j$ for input utterance $(X)$ is

$$C(t_j|X) = \sum_{i=1}^{i \leq N} p(s_i|X)c_{SVM}(t_j|W_i);$$

$W_i$       feature vector representation of $s_i$;

$c_{SVM}(t_j|W_i)$ classification score of topic $t_j$ for feature vector $W_i$;

$p(s_i|X)$       posterior probability of $i$th recognition hypothesis $s_i$ by ASR      (3)

where the posterior probability $p(s_i|X)$ is calculated from the combined ASR probabilities (language model probability and

$T \leftarrow \{t_1, t_2, \ldots, t_M\}, k \leftarrow 1$
**do**
    **select** $t_i, t_j$ such that $dist(t_i, t_j)$ is minimum $(i \neq j)$
    **merge** $t_i$ and $t_j$ to create *meta-topic* $t_k$
    **remove** $t_i$ and $t_j$ from $T$
    **add** *meta-topic* $t_k$ to active set $T$
    $k \leftarrow k + 1$
**while** $dist(t_i, t_j) < thres_{hierarchy}$

acoustic model probability) in the N-best list as described in [15]

$$p(s_i|X) = \frac{e^{\gamma \cdot \text{logscore}_i}}{\sum_{j=1}^{j \leq N} e^{\gamma \cdot \text{logscore}_j}};$$

$\text{logscore}_i$    log-scale ASR score;

$\gamma$          smoothing factor $(0 < \gamma < 1)$.    (4)

The smoothing factor $\gamma$ is introduced to give an adequate distribution of confidence measures and is dependent on the acoustic and language models applied during recognition.

### D. Topic Clustering for Robust Topic Classification

In spontaneous dialogue, utterances are typically short, ungrammatical and often contain elliptical and anaphoric elements. Thus, the relationship between utterances and individual topic-classes tend to be indistinct. To improve the robustness of topic classification, we investigate a topic clustering approach where a set of *meta-topic* classes are generated automatically to provide coverage over closely related and confusable topic classes. An overview of the clustering procedure, which we first proposed in [27], is shown in Table II.

*Meta-topics* are generated automatically by performing agglomerative clustering on the individual topic classes. Clustering involves iteratively selecting the closest topic pairs and merging them until the distances between all topics is greater than some pre-defined threshold. The distance measure applied during clustering $dist(t_i, t_j)$ is based on the confusability between topics and is defined as the average distance between topic $t_i$'s training data $(T_i)$ and topic $t_j$'s SVM hyperplane and vice versa (5)

$$dist(t_i, t_j)$$
$$= \left\| \underset{W \in T_i}{\text{average}}\, dist_\perp(W, H_j) - \underset{W \in T_j}{\text{average}}\, dist_\perp(W, H_j) \right\|$$
$$+ \left\| \underset{W \in T_j}{\text{average}}\, dist_\perp(W, H_i) - \underset{W \in T_i}{\text{average}}\, dist_\perp(W, H_i) \right\|$$

$T_i$          set of training sentences of topic class $t_i$;

$dist_\perp(W, t_j)$ perpendicular distance from sentence $W$ to SVM hyperplane of topic $t_j(H_j)$.   (5)

The resulting dendrogram for an evaluation task is shown in Fig. 4. In this example, six *meta-topic* clusters were generated $\{\#1, \ldots, \#6\}$. The lowest layer of the structure
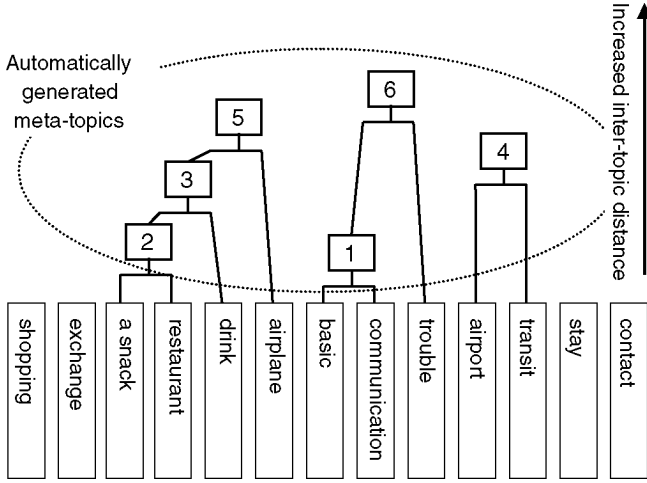
Fig. 4. Generated *meta-topic* clusters for evaluation task.

corresponds to the original individual topic classes and classes higher in the structure correspond to the *meta-topics* that provide coverage over multiple topic-classes. Topic classification models are trained for all topic classes within this hierarchy, and these models are applied during classification. When *meta-topics* are incorporated into topic classification, an extended topic confidence vector is generated $(C(t_1|X), \ldots, C(t_M|X), C(t_{M+1}|X), \ldots, C(t_{M+M_{\mathrm{meta}}}|X))$, where $(C(t_{M+1}|X), \ldots, C(t_{M+M_{\mathrm{meta}}}|X))$ corresponds to the confidence scores of the *meta-topic* classes and $M_{\mathrm{meta}}$ is the number of *meta-topic* classes.

Incorporating *meta-topics* into the proposed framework enables the system to assign an utterance to a cluster of topics even when the exact individual topic cannot be identified. For example, suppose an utterance "*excuse me, I would like to get to a hotel in town, what would be the best way to get there?*" and the clustering structure shown in Fig. 4. Applying only individual topic classes, two classes "*airport*" and "*transit*", have the highest confidence scores of 33% and 2%, respectively. However, when topic clustering is applied, *meta-topic* #4, has the highest confidence score (92%), indicating that the utterance is in-domain.

## IV. IN-DOMAIN VERIFICATION

In the final stage of OOD detection (Fig. 2, stage 3) an in-domain verification model $V(X)$ is applied to the topic confidence vector generated during topic classification. In this paper, a linear discriminant model, shown in (6), is adopted

$$V(X) = \begin{cases} 1, & \text{if } \sum_{j=1}^{j \leq M} \lambda_j C(t_j|X) \geq \varphi \quad \text{(in-domain)} \\ 0, & \text{otherwise.} \quad \text{(OOD)} \end{cases}$$

$C(t_j|X)$ classification score of topic $t_j$
    for input utterance $X$;
$M$    number of topic-classes.     (6)

Linear discriminant weights $\{\lambda_1, \ldots, \lambda_M\}$ are applied to the confidence scores of topic classification $(C(t_1|X), \ldots, C(t_M|X))$, and the weighted sum is compared

### TABLE III
VERIFIER TRAINING BASED ON DELETED INTERPOLATION

**for each** topic $t_i$ in $\{t_1, \ldots, t_M\}$
  set topic $t_i$ as temporary OOD
  set remaining topic classes as in-domain
  calculate $\{\lambda_1, \ldots, \lambda_M\}$ using GPD ($\lambda_i$ excluded)
average $\{\lambda_1, \ldots, \lambda_M\}$ over all iterations

to a threshold ($\varphi$). If the verification score is greater than this threshold, the utterance is classified as in-domain. Otherwise, it is classified as OOD.

If both in-domain and OOD training data are available, the discriminant weights $\{\lambda_1, \ldots, \lambda_M\}$ can be trained to minimize classification errors using discriminative training. However, it is often the case that task-dependent OOD training data is not available. To overcome this problem, we introduce a novel scheme to train the verification model using only in-domain data. This is described in the following subsections.

### A. Verifier Training Based on Deleted Interpolation of Topics

The proposed training scheme combines deleted interpolation and minimum-classification-error training using the GPD (gradient probabilistic descent) algorithm [28]. An overview of the proposed method is given in Table III. During each training iteration, a single topic class $t_i$ is set to be temporarily OOD, and the corresponding vector component $C(t_i|X)$ is removed from the verification model. The discriminant weights of the remaining topic classifiers $\{\lambda_j, 1 \leq j \leq M, j \neq i\}$ are estimated using GPD. Upon completion of all iterations, the final model weights $\{\lambda_1, \ldots, \lambda_M\}$ are calculated by averaging over all interpolation steps.

During each iteration, a subset of the classifier coefficients $\{\lambda_j, 1 \leq j \leq M, j \neq i\}$ are discriminatively trained using the GPD algorithm to minimize classification errors. In this step, the training set of temporary OOD data $T_i$ is used as negative training examples, and a balanced set of the remaining topic classes is used as positive (in-domain) examples. During hypothesis testing, the discriminative functions for the null hypothesis ($H_0$: in-domain), and the alternative hypothesis ($H_1$: out-of-domain), are defined as follows:

$$H_0 : g_0(X) = \sum_{j=1, j \neq i}^{j \leq M} \lambda_{j_0} C(t_{j_0}|X) \quad \text{[In-domain]}$$

$$H_1 : g_1(X) = \sum_{j=1, j \neq i}^{j \leq M} \lambda_{j_1} C(t_{j_1}|X) \quad \text{[Out-of-domain]}$$

and the misclassification measure is

$$d(X) = -g_0(X) + g_1(X)$$
$$= \sum_{j=1, j \neq i}^{j \leq M} \theta_j C(t_j|X) \qquad (7)$$
$$\theta_j = -\lambda_{j_0} + \lambda_{j_1}. \qquad (8)$$

By minimizing the loss function (9), which is defined as the sigmoid function ($sigmoid[]$) of the misclassification measure,

we obtain the optimal set of classification coefficients $\theta_j, 1 \leq j \leq M, j \neq i$

$$sigmoid\,[d(X)] = \frac{1}{1 + e^{(\alpha \cdot d(X) + \beta)}}. \tag{9}$$

More specifically, its derivative $\theta_j$ is adjusted by $\Delta\theta_j$ according to (10) for the null hypothesis and (11) for the alternative hypothesis

$$
\begin{aligned}
H_0 : \Delta\theta_j &= -\varepsilon\nabla sigmoid\,[d(X)] \\
&= -\varepsilon sigmoid'\,[d(X)]\frac{\partial d(X)}{\partial\theta_{j_0}} \\
&= \varepsilon\frac{\alpha \cdot e^{(\alpha \cdot d(X)+\beta)}}{\{1 + e^{(\alpha \cdot d(X)+\beta)}\}^2}C(t_j|X)
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
H_1 : \Delta\theta_j &= -\varepsilon\nabla sigmoid\,[d(X)] \\
&= -\varepsilon sigmoid'\,[d(X)]\frac{\partial d(X)}{\partial\theta_{j_1}} \\
&= -\varepsilon\frac{\alpha \cdot e^{(\alpha \cdot d(X)+\beta)}}{\{1 + e^{(\alpha \cdot d(X)+\beta)}\}^2}C(t_j|X).
\end{aligned} \tag{11}
$$

During training, the discriminant weight for $t_j$, $\lambda_j$ $(\Delta\theta_j)$, is influenced predominantly by training examples with high $C(t_j|X)$; either actual training data from $t_j$ (10), or temporary OOD data from another topic that receives high $C(t_j|X)$ (11). The proposed scheme is thus designed to find the optimal weights for individual topics based on their confusability.

### B. Verification Modeling for Topic Clustering

When *meta-topics* (generated using the topic clustering scheme in Section III-D) are incorporated, the total number of topic classification models is increased. The verification model must be updated to match this. In this case, the verifier consists of $M + M_{\mathrm{meta}}$ linear discriminant weights $\{\lambda_1, \ldots, \lambda_M, \lambda_{M+1}, \ldots, \lambda_{M+M_{\mathrm{meta}}}\}$, where $M$ is the number of individual topics, $M_{\mathrm{meta}}$ is the number of *meta-topics*, and $\{\lambda_{M+1}, \ldots, \lambda_{M+M_{\mathrm{meta}}}\}$ are the discriminate weights of the *meta-topics* classifiers. The discriminative weights are trained using the scheme described in the previous sub-section, however, during training *meta-topic* classifiers that are parents of the current temporary OOD topic $(t_j)$ are also removed for that iteration step.

### V. Incorporating Dialogue Context

When applying OOD detection to spoken dialogue tasks that are completed via multiple utterances, an OOD 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 $V(X^{[1,\ldots,n]})$ 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 approaches are described in the following subsections.

### A. Word Vector-Level Combination (WRD)

A common approach used to combine multiple utterances is to concatenate the word sequences $(X_1, \ldots, X_n)$ and generate a single word vector $W^{[1,\ldots,n]}$ (12). Topic classification can then be applied to this vector. The resulting scores are used for in-domain verification (13)

$$W^{[1,\ldots,n]} = \sum_{j=1}^{j \leq n} W^j \tag{12}$$

$$
\begin{aligned}
&V_{WRD}\left(X^{[1,\ldots,n]}\right) \\
&= \begin{cases} 1, & \text{if } \sum_{j=1}^{j \leq M} \lambda_j c_{SVM}\left(t_j|W^{[1,\ldots,n]}\right) \geq \varphi \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{13}
$$

### B. 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 (14), generating a single topic classification vector. In-domain verification is then applied to this vector (15).

$$C_{\mathrm{avg}}(t_i|X^1, \ldots, X^n) = \frac{1}{n}\sum_{j=1}^{j \leq n} C(t_i|X^j) \tag{14}$$

$$
\begin{aligned}
&V_{TOP}\left(X^{[1,\ldots,n]}\right) \\
&= \begin{cases} 1, & \text{if } \sum_{j=1}^{j \leq M} \lambda_j C_{\mathrm{avg}}(t_j|X^1, \ldots, X^n) \geq \varphi \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{15}
$$

### C. Verification-Level Combination (VER)

In this method, topic classification and in-domain verification are applied independently for each input utterance. The final decision is made by averaging over the individual verification scores (16)

$$
\begin{aligned}
&V_{VER}\left(X^{[1,\ldots,n]}\right) \\
&= \begin{cases} 1, & \text{if } \frac{1}{n}\sum_{i=1}^{i \leq n}\left(\sum_{j=1}^{j \leq M} \lambda_j C\left(t_j|X_j\right)\right) \geq \varphi \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{16}
$$

### VI. Experimental Evaluation

The performance of the proposed OOD detection framework was evaluated on two tasks: a "*phrasebook*" system [10], which performs Japanese-to-English translation for individual read-style spoken utterances, and a "*Machine aided dialogue*" (*MAD*) system [29], which enables Japanese and English speakers to communicate via spoken dialogue using parallel Japanese-to-English and English-to-Japanese translation systems. Both systems were developed specifically to operate within an "*overseas travel*" domain which covers utterances users are likely to use or encounter while traveling overseas. Utterances which fall outside this application domain cannot be correctly translated by the back-end machine translation system, and are defined as

TABLE IV
ATR BASIC TRAVEL EXPRESSION CORPUS (BTEC)

| | | | |
|---|---|---|---|
| **Domain**: Overseas travel | | | |
| **In-domain**: | 11 topics (*transit, accommodation, ...*) | | |
| **Set as OOD**: | 1 topic (*shopping*) | | |
| **Training set**: 11 topics, 149540 sentences (In-domain data only) | | | |
| **Lexicon size**: 17000 words | | | |
| **Test-set**: | *"misc"* | **In-Domain**: | 1852 utterances |
| | | **OOD**: | 398 utterances |
| | *"shopping"* | **In-Domain**: | 1852 utterances |
| | | **OOD**: | 138 utterances |

being OOD. As the application domain is complex, OOD detection is required to realize an effective user interface.

### A. Phrasebook Task

For the first experimental evaluation, OOD detection was applied to the "*phrasebook*" task of the speech-based Japanese-to-English translation system. For this evaluation, two sets of OOD utterances were evaluated. The first set "*misc*" contains utterances that are not related to the application domain, but are likely to be encountered by the system, and the second set "*shopping*" simulates the case when an in-domain topic is not covered by the application. For the second evaluation set, we set one topic class, in this case the topic "*shopping*" to be out-of-application of the back-end system[1] (e.g. we assume the back-end MT system has not been implemented to translate "shopping" utterances) and excluded this data from the training corpus.

The ATR basic travel expressions corpus (ATR-BTEC) [30], described in Table IV, was used for evaluation. The training set consists of 149 540 sentences made up from 11 in-domain topic classes. This data was used to train the language model applied during speech recognition as well as the classification and verification models required for OOD detection. The "*misc*" and "*shopping*" test-sets consist of 2250 and 1990 utterances, respectively. Example test-set utterances are given in Table V.

System performance was evaluated on the ability to discriminate between in-domain and OOD utterances. The following evaluation measures were used.

**FRR** (False Rejection Rate): Percentage of in-domain utterances classified as OOD.

**FAR** (False Acceptance Rate): Percentage of OOD utterances classified as in-domain.

**EER** (Equal Error Rate): Error rate at an operating point where FRR and FAR are equal.

Based on these evaluation measures, we evaluated the following aspects of the proposed OOD detection framework: the discriminative ability of various feature sets used during topic classification, the performance of the proposed deleted interpolation training scheme, and the robustness of the framework against speech recognition errors.

*1) Effect of Topic Classification Features:* First, the discriminative ability of the topic classification feature sets described in Section III-A were investigated. Initially, SVM-based topic classification models were trained for each feature set using the ATR-BTEC training set. A closed evaluation was then performed using the correct transcriptions of the "*misc*" test-set.

Topic classification confidence scores were first calculated for the in-domain and OOD sets by applying the above SVM models, and this data was then used to train the in-domain verification model. During training, in-domain data were used as positive training examples and OOD data were used as negative training examples. The performance of each feature set was then evaluated by applying this closed-model to the same confidence vectors used for training. The performance for each feature set is shown in Table VI in terms of the topic classification accuracy of the in-domain utterances (column 4) and OOD detection EER (column 5).

When word baseform features were used, a topic classification accuracy of 87.4% and an OOD detection EER of 20.6% were gained. The inclusion of context-based features, consisting of word-pairs (2-gram) and word-triplets (3-gram), significantly improved OOD detection accuracy and improved topic classification performance. A minimum OOD detection EER of 10.8% was obtained when word+POS tokens and 1, 2, and 3-gram features were incorporated. Hereafter, this feature set is adopted for OOD detection.

Incorporating context-based features improved OOD detection performance significantly, but the improvement in topic classification accuracy was limited. These features are useful for detecting erroneous word sequences within an utterance, and are thus useful for OOD detection, however, such information is irrelevant for topic classification of in-domain topics. Incorporating n-gram information during classification increases the size of the classification space significantly. However, SVM considers only the discriminative features in the support vectors. For example, the classifier for the topic "*airplane*" uses only 30% of the total available features for classification. In earlier work [31], we investigated the performance of two other topic-classification schemes, topic-dependent N-grams and LSA (latent semantic analysis), and found that SVM significantly outperformed both these approaches.

*2) Performance of Deleted Interpolation-Based Training:* Next, the performance of the proposed deleted interpolation-based training scheme described in Section IV-A was evaluated. Again, the correct transcriptions of the "*misc*" test-set were used. We compared the OOD detection performance of three systems: a system trained using the proposed deleted interpolation-based scheme (*proposed*), a reference method (as adopted in the previous experiment) in which the in-domain verification model was trained using both in-domain and OOD data from the test-set (*closed-model*), and a baseline system (*baseline*). In the baseline system, the maximum score from topic classification SVMs was compared to a threshold $\varphi$. An utterance was classified as in-domain if the resulting score was greater than the threshold $(\max_i C(t_i|X) > \varphi)$, and OOD otherwise. The error rate (FRR vs. FAR) curves of the three systems are obtained by iteratively increasing the verification threshold $\varphi$. These curves are plotted in Fig. 5.

The baseline system has an EER of 18.9%. The proposed method provided a significant reduction in detection errors compared to this baseline obtaining an EER of 10.0% (for the same FRR (18.9%) as the baseline, an FAR of 3.5% was gained). Furthermore, the proposed method achieved comparable performance to the closed evaluation case (*closed-model*). This shows

---

[1]This is not related to the temporary OOD topic in deleted interpolation

TABLE V
EXAMPLE UTTERANCES FOR IN-DOMAIN AND OOD TOPIC CLASSES

| Topic Class | No. of Training Sentences | | Example Sentences |
|---|---|---|---|
| **In-Domain Topic Classes** | | | |
| accommodation | 16473 | (11%) | "I asked for a room with a shower.", "What's the price, excluding meals?" |
| airplane | 7117 | (5%) | "Which channel is the film on?", "What kind of drinks do you have?" |
| airport | 10101 | (7%) | "Please show me your passport.", "Where can I change travelers cheques?" |
| basic | 21880 | (15%) | "Sorry.", "Hi.", "What does that mean?" |
| communication | 11613 | (8%) | "Are you from New Zealand?", "May I use your bathroom?" |
| contact | 7514 | (5%) | "Extension two thirty four please.", "Call to Japan please.", "Sorry wrong number." |
| eat&drink | 18746 | (13%) | "Could we have an ashtray please.", "I'd like some bread." |
| exchange | 2151 | (1%) | "Cash this please.", "How would you like it?", "Traveler's checks okay?" |
| sightseeing | 14581 | (10%) | "Do you know where the baseball stadium is?", "Could you take a photo of us, please?" |
| transit | 17027 | (11%) | "Is this the right platform for the train to Minneapolis?" |
| trouble | 22337 | (15%) | "Can you send a mechanic please?", "There's something wrong with the clutch." |
| **Out-of-Domain Data** | | | |
| misc | - | | Utterances not related to "overseas travel" domain; e.g. beauty, business, study, etc. |
| shopping | - | | "Do you have any bags around $200.", "I am a size eleven.", "This is too large for me." |

TABLE VI
COMPARISON OF TOPIC CLASSIFICATION FEATURE SETS

| Token Set | Feature Set | Number of Features | Topic Classification Accuracy | OOD Detection EER |
|---|---|---|---|---|
| baseform | 1-gram | 8771 | 87.4% | 20.6% |
| word | 1-gram | 9899 | 87.7% | 20.3% |
| word+POS | 1-gram | 10006 | 87.0% | 20.2% |
| word+POS | 1,2-gram | 40754 | 90.6% | 15.3% |
| word+POS | 1,2,3-gram | 73065 | 90.2% | 10.8% |

TABLE VII
SPEECH RECOGNITION PERFORMANCE FOR PHRASEBOOK DATA

| | # Utt. | WER(%) | SER(%) | OOV(%) |
|---|---|---|---|---|
| **In-Domain** | 1852 | 7.26 | 22.4 | 0.71 |
| **Out-of-Domain Data** | | | | |
| *misc* | 398 | 28.8 | 78.9 | 5.52 |
| *shopping* | 138 | 12.49 | 45.3 | 2.56 |

**WER**: Word Error Rate  **SER**: Sentence Error Rate
**OOV**: Out of Vocabulary Rate



Fig. 5.  OOD detection performance on correct transcriptions ("*misc*").
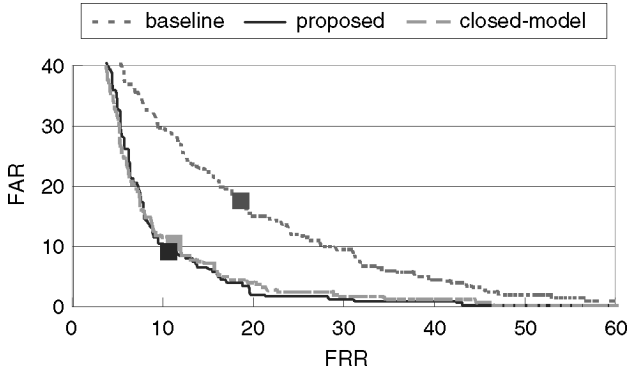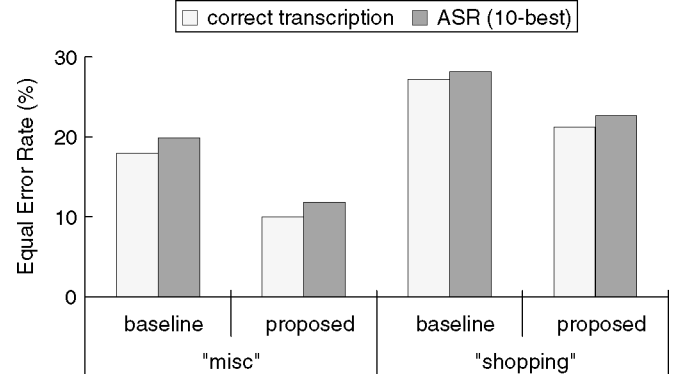


Fig. 6.  OOD detection performance on ASR results (phrasebook).

that the deleted interpolation-based training realizes a near optimal verification model even in the absence of OOD data.

*3) Application to ASR Results:* The performance of the proposed system was then evaluated on the automatic speech recognition (ASR) results of the "*misc*" and "*shopping*" test-sets. ASR was performed with our Julius recognition engine [32]. For acoustic analysis, 12-dimensional mel frequency cepstral coefficients (MFCC), energy, and their first and second-order derivatives were computed. The acoustic model was a triphone HMM with 1841 shared states and 23 Gaussian mixture components set up for 26 phones. A word trigram language model trained on the entire in-domain training set (149 540 sentences) was applied. The ASR performance for the in-domain and OOD test-sets is shown in Table VII.

Topic classification confidence vectors were generated from the 10-best recognition hypotheses using the method described in Section III-C, and a smoothing factor of $\gamma = 0.3$ (4) was

applied. The EERs of the baseline and proposed systems when applied to the correct transcriptions and ASR results are shown in Fig. 6. For the "*misc*" test-set, an EER of 11.8% was gained with the proposed system. This is an absolute increase of 1.8 points compared to the case for the correct transcriptions, and a reduction in EER of 8.1 points compared to the baseline system. This demonstrates that the proposed approach is robust against recognition errors. Using the 10-best ASR hypotheses during topic classification improved accuracy by 1.3 points compared to using the 1-best hypothesis alone (EER = 13.1%).

A similar result was observed for the "*shopping*" test-set. For the ASR case, EERs of 29.8% and 22.7% were gained for the baseline and proposed systems, respectively. The overall detection accuracy for this test-set, however, is significantly lower than that of the "*misc*" set. This indicates that OOD utterances which are related to the application domain of the system, but are not covered by the back-end application, are more difficult

TABLE VIII
TRAINING CORPUS FOR MAD TASK

| | |
|---|---|
| **Domain**: | Overseas travel |
| **Languages**: | English, Japanese |
| **Training Set**: | 14 topics (*accommodation, shopping...*), 400k sentences |
| **Lexicon Size**: | 10k/20k (English/Japanese respectively) |
| **test-set**: | 305 dialogue sessions, 2964/3178 utterances (English/Japanese respectively) |

TABLE IX
EVALUATION OF TOPIC CLUSTERING (MAD; TRANSCRIPTION)

| Initiating speaker | OOD Topic | No. Sessions OOD | No. Sessions ID | OOD detection accuracy (EER%) topic | OOD detection accuracy (EER%) cluster |
|---|---|---|---|---|---|
| Japanese | accommodation | 44 | 111 | 27.6 | 20.6 |
| | shopping | 22 | 132 | 23.1 | 13.6 |
| | sightseeing | 20 | 134 | 28.4 | 24.8 |
| | **TOTAL** | **86** | **377** | **26.4** | **19.7** |
| English | accommodation | 37 | 113 | 15.6 | 11.2 |
| | shopping | 11 | 140 | 13.0 | 13.0 |
| | sightseeing | 20 | 131 | 23.2 | 15.1 |
| | **TOTAL** | **68** | **384** | **17.3** | **13.1** |

**topic**: classifiers applied for original topics only
**cluster**: *meta-topic* clusters incorporated

to detect than OOD utterances which are outside the application domain. For both test sets the difference between the baseline and proposed system was statistically significant at $p < 0.01$ (McNemar test).

### B. Machine-Aided Dialogue (MAD) Task

For the second experimental evaluation, we applied OOD detection to the ATR machine-aided dialogue system [29], a bi-directional speech-to-speech translation system. The system consists of two speech-to-speech translation systems, an English-to-Japanese system and a Japanese-to-English system, operating in parallel. OOD detection was integrated into this system independently for each language side. The test-set used for evaluation consists of 305 dialogue sessions between native English and Japanese speakers. Each session was conducted based on a pre-defined dialogue scene, such as, *asking for directions*, *partaking in a meal*, *or finding one's lost luggage*. In this task, speech is much more spontaneous than in the previous "*phrasebook*" evaluation.

Evaluation was performed for three test scenarios. In each scenario, one topic from the corpus was set as OOD of the system (Table IX column 2), and this topic was excluded from the training set. The language model for speech recognition and OOD detection modules were then trained on the remaining in-domain topic data. During training, an extended version of the ATR-BTEC corpus consisting of 400k sentences and 14 topics was used (Table VIII). In this set of experiments, FAR and FRR are evaluated for entire dialogue sessions, rather than utterances as in the "*phrasebook*" experiment.

*1) Effect of Topic Clustering:* First, we evaluated the effectiveness of topic-clustering, as proposed in Section III-D. In this experiment, OOD detection was applied to the correct transcriptions of the initial ($n = 1$) utterance of each dialogue. The OOD detection performance for the three test scenarios when only the original topic classifiers were applied (*topic*) and when *meta-topics* (generated during topic clustering) were included (*cluster*) are shown in Table IX, columns 5 and 6, respectively. The threshold applied during meta-topic clustering ($\text{thres}_{\text{hierarchy}}$, Table II) was empirically chosen, but was consistent for all test scenarios.

Topic clustering provided a total reduction in OOD detection EER of 6.7 points (from 26.4% to 19.7%) for the Japanese side and 4.2 points (from 17.3% to 13.1%) for the English side. 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. For read-style speech of the "*phrasebook*" task (Section VI-A), on the other hand, the effectiveness of topic clustering was limited.

As input utterances are typically grammatically correct, complete and limited to a single topic, improved topic classification robustness was not required for this task.

*2) Effect of Dialogue Context:* Next, we investigated the system performance when dialogue context was incorporated into the OOD detection framework. We compared three methods to combine multiple utterances as described in Section V. The performance when applied to the correct transcriptions is shown in Fig. 7. Each method was evaluated when applied to the first $n$ utterances of the dialogue, for $n = 1,2,3$. Topic clustering was not incorporated in this experiment. Combining utterances at the topic classification-level (TOP) provided the best performance with a reduction in EER of 4.8 points (from 26.4% to 21.5%) for the Japanese side ($n = 3$), and 1.9 points (from 17.3% to 15.4%) for the English side ($n = 3$). This improvement, however, is smaller than that gained by topic-clustering, suggesting that the initial utterance is typically the most relevant for OOD detection.

Utterance combination at the word vector-level (WRD) and in-domain verification-level (VER) resulted in poorer detection accuracy. At the word-vector-level, a shift in topic within the dialogue cannot be represented adequately by a single word-vector. At the verification-level, the dynamic range of scores is large, thus averaging is not effective as it tends to be affected by outliers.

*3) Overall System Performance:* Finally, topic clustering and utterance combination (at the topic classification-level) were combined and the system was evaluated on both the speech recognition (ASR) results as well as the correct transcriptions. Speech recognition was performed using a setup similar to that in the "*phrasebook*" task (Section VI-A-3), and language models trained using only in-domain data were applied. The average WER of the Japanese dialogue side for the in-domain and OOD sets is shown in Table X. As the English ASR is still under development, we did not integrate it in this work.

The OOD detection performance for the original framework, and when topic clustering and dialogue context were incorporated, is shown in Table XI. For the transcription case a significant reduction in OOD detection errors was gained by combining these two approaches. Individually, topic clustering and utterance combination provided a reduction in EER of 6.7 and 4.8 points, respectively. When combined, a total reduction in EER of 9.0 points (from 26.4% to 17.3%) was gained for the

Fig. 7. Evaluation of utterance combination (MAD; transcription). (a) Japanese side; (b) English side.
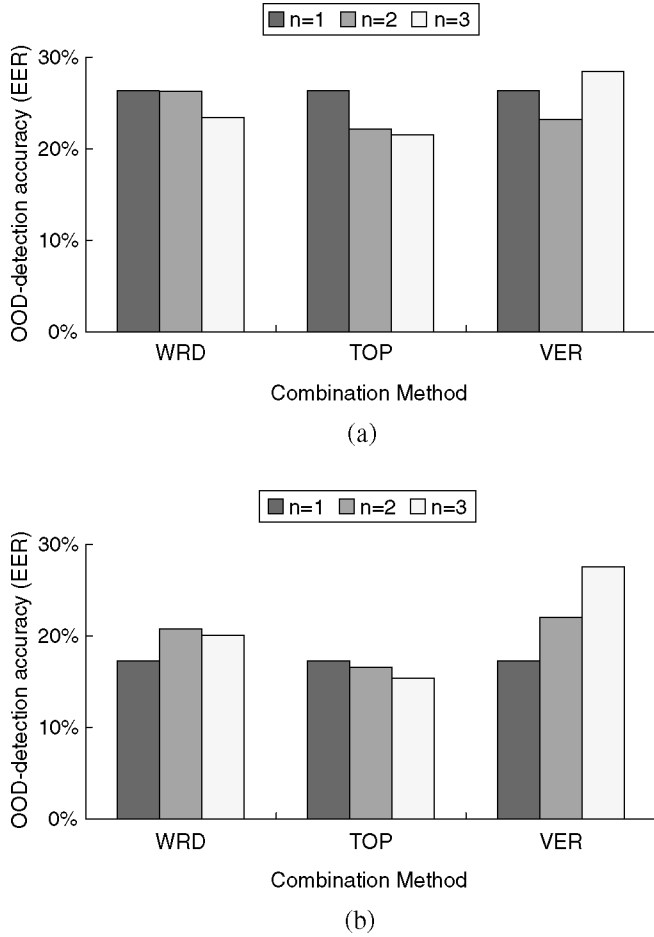
WRD: word vector-level combination
TOP: topic classification-level combination
VER: verification-leve lcombination

TABLE X
SPEECH RECOGNITION PERFORMANCE FOR MAD DATA

| | In-domain | | Out-of-domain | |
|---|---|---|---|---|
| **Language** | **WER** | **SER** | **WER** | **SER** |
| Japanese | 11.8% | 40.2% | 15.5% | 55.0% |

WER: word error rate    SER: sentence error rate

TABLE XI
OOD DETECTION PERFORMANCE ON TRANSCRIPTIONS
AND ASR RESULTS (MAD)

| System | OOD detection EER (relative reduction) | |
|---|---|---|
| | Transcription | ASR |
| Original | 26.4% | 28.3% |
| Topic Clustering | 19.7% (25%) | 20.9% (26%) |
| Topic Clustering + Dialogue Context ($n$=3) | 17.3% (34%) | 20.4% (28%) |

EER by 7.4% to 20.9%, which is similar to that gained for the transcript case (EER $= 19.7\%$). This indicates that the proposed framework incorporating topic clustering is robust against ASR errors. The effectiveness of incorporating dialogue context via utterance combination, however, is reduced. For the ASR case an EER of 20.4% was gained. When applied to ASR results, recognition errors may lead to erroneous topic-classification, and as the number of utterances $n$ is increased from $n = 1$ to $n = 3$, the likelihood of topic-classification errors also increases. This limits the effectiveness of this approach especially when ASR errors are common. To improve system robustness, utterances with low ASR confidence (those more likely to be affected by ASR errors) should be removed from consideration during utterance combination.

Finally, we tested the system performance on a currently available set of real OOD dialogues. We collected a set of 139 in-domain and 12 OOD dialogues (not related to the application domain "overseas travel"). OOD detection, incorporating topic clustering and dialogue context ($n = 3$), was then applied to the ASR results of these dialogues. The proposed system successfully rejected all OOD dialogues within the first 3 utterances, while accepting 86% of in-domain dialogues. In future work we intend to perform a full evaluation using real-world data collected via a deployed speech-to-speech translation system.

## VII. CONCLUSION

In this paper, we have proposed a novel OOD (out-of-domain) detection framework which uses topic classification confidence for in-domain verification. We also introduced a novel verification model training scheme, based on deleted interpolation of the in-domain data and minimum classification error training. This scheme enables the OOD detection module to be realized using only in-domain training data. In the "*phrasebook*" task, OOD detection errors were reduced by 8.1 points (40% relative) compared to a baseline system based on the maximum topic classification score. Furthermore, similar performance to an equivalent system trained on both in-domain and OOD data was gained while requiring no OOD data during training. We also compared various feature sets to be applied during topic classification, and observed that the incorporation of word, word-pair, and word-triplet features was important for effective performance. We also applied OOD detection to the "*machine aided dialogue*" system. For this more spontaneous task, incorporating *meta-topics* during topic classification significantly improved OOD detection performance. Incorporating dialogue context into the proposed framework also provided some improvement.

In the above experimental evaluations, we evaluated the proposed OOD detection framework with speech-to-speech translation systems. However, this framework is not limited to this task and can easily be incorporated into other spoken language systems, for example, call-routing and spoken dialogue systems. Before implementing the proposed framework, however, an adequate set of pre-defined topic classes is required. These can be defined by call destinations in a call-routing system, or sub-domains in a spoken dialogue system. Automatic generation of these classes via sentence clustering, as described in [33],

$n = 3$ case. This shows the effectiveness of combining the two proposed approaches.

For the ASR case, a similar reduction in EER was gained (EER reduced by 7.9 points from 28.3% to 20.4%). Topic clustering provided the most significant improvement, reducing

should also be explored. In future work, we also intend to investigate approaches to improve discrimination of OOD utterances and erroneously recognized in-domain utterances.

## REFERENCES

[1] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 1, pp. 1–8, Jan. 2000.

[2] S. Seneff and J. Polifroni, "A new restaurant guide conversational system: Issues in rapid prototyping for specialized domains," in *Proc. ICSLP*, 1996, vol. 2, pp. 665–668.

[3] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken-language understanding in the MIT Voyager system," *Speech Commun.*, vol. 17, pp. 1–18, 1995.

[4] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker, "DARPA Communicator dialog travel planning systems: The June 2000 data collection," in *Proc. Eurospeech*, 2001, pp. 1371–1374.

[5] A. Gorin, G. Riccardi, and J. Wright, "How may I help you?," *Speech Commun.*, vol. 23, pp. 113–127, 1997.

[6] P. Durston, M. Farrell, D. Attwater, J. Allen, H.-K. J. Kuo, M. Afify, E. Fosler-Lussier, and C.-H. Lee, "OASIS natural language call steering trial," in *Proc. Eurospeech*, 2001, vol. 2, pp. 1323–1326.

[7] C.-H. Lee, B. Carpenter, W. Chou, J. Chu-Carroll, W. Reichl, A. Saad, and Q. Zhou, "On natural language call routing," *Speech Commun.*, vol. 31, no. 4, pp. 309–320, 2000.

[8] E. W. Wahlster, *VERBMOBIL: Foundations of Speech-to-Speech Translation*. Berlin, Germany: Springer, 2000.

[9] A. Lavie, F. Metze, R. Cattoni, E. Costantini, S. Burger, D. Gates, C. Langley, K. Laskowski, L. Levin, K. Peterson, T. Schultz, A. Waibel, D. Wallace, J. McDonough, H. Soltau, G. Lazzari, N. Mana, F. Pianesi, E. Pianta, L. Besacier, H. Blanchon, D. Vaufreydaz, and L. Taddei, "A multi-perspective evaluation of the NESPOLE! speech-to-speech translation system," in *Proc. ACL, Workshop on Speech-to-Speech Translation*, 2002, pp. 121–128.

[10] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, "A Japanese-to-English speech translation system: ATR-MATRIX," in *Proc. ICSLP*, 1998, pp. 957–960.

[11] T. Hanzen, J. Polifroni, and S. Seneff, "Recognition confidence scoring for use in speech understanding systems," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 49–67, 2002.

[12] C. Ma, M. Randolph, and J. Drish, "A support vector machines-based rejection technique for speech recognition," in *Proc. ICASSP*, 2001, vol. 1, pp. 381–384.

[13] R. Sukkar and C. Lee, "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 6, pp. 420–429, Nov. 1996.

[14] R. San-Segundo, B. Pellom, and W. Ward, "Confidence measures for dialogue management in the CU communicator system," in *Proc. ICASSP*, 2000, vol. 2, pp. 1237–1240.

[15] K. Komatani and T. Kawahara, "Generating effective confirmation and guidance using two-level confidence measures for dialogue systems," in *Proc. ICSLP*, 2000, pp. 648–651.

[16] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inform. Sci.*, vol. 41, pp. 391–407, 1990.

[17] S. Cox and B. Shahshahani, "A comparison of some different techniques for vector based call-routing," in *Proc. Eurospeech*, 2001, vol. 4, pp. 2337–2340.

[18] T. Joachims, "Text categorization with support vector machines," in *Proc. Eur. Conf. Machine Learning*, 1998, pp. 137–147.

[19] P. Haffner, G. Tur, and J. Wright, "Optimizing SVMs for complex call classification," in *Proc. ICASSP*, 2003, vol. 1, pp. 632–635.

[20] Y. Wang, A. Acero, C. Chelba, B. Frey, and L. Wong, "Combination of statistical and rule-based approaches for spoken language understanding," in *Proc. ICSLP*, 2002, pp. 609–612.

[21] G. Tur and D. Hakkani-Tur, "Exploiting unlabeled utterances for spoken language understanding," in *Proc. Eurospeech*, 2003, pp. 2793–2796.

[22] I. Zitouni, H.-K. J. Kuo, and C.-H. Lee, "Combination of boosting and discriminative training for natural language call steering systems," in *Proc. ICASSP*, 2002, vol. 1, pp. 25–28.

[23] P. Liu, H. Jiang, and I. Zitouni, "Discriminative training of naive bayes classifiers for natural language call routing," in *Proc. ICSLP*, 2004, pp. 1589–1592.

[24] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers," in *Proc. Conf. Artificial Intelligence*, 1992, pp. 223–228.

[25] H.-K. J. Kuo and C.-H. Lee, "Discriminative training of natural language call routers," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 1, pp. 24–35, Jan. 2003.

[26] I. Lane, T. Kawahara, T. Matsui, and S. Nakamura, "Dialogue speech recognition by combining hierarchical topic classification and Language model switching," *IEICE Trans. Inform. Syst.*, vol. E88-D, no. 3, pp. 446–454, 2005.

[27] I. Lane and T. Kawahara, "Incorporating dialogue context and topic clustering in out-of-domain detection," in *Proc. ICASSP*, 2005, vol. 1, pp. 1045–1048.

[28] S. Katagiri, C.-H. Lee, and B.-H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method," in *Proc. IEEE Int. Workshop on Neural Networks for Signal Process.*, 1991, pp. 299–308.

[29] T. Takezawa, A. Nishino, K. Takashima, T. Matsui, and G. Kikui, "An experimental system for collecting machine-translation aided dialogues," in *Forum on Information Technology (Japanese)*, 2003, vol. 2, pp. 161–162.

[30] 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. Int. Conf. on Language Resources and Evaluation*, 2002, pp. 147–152.

[31] I. Lane, T. Kawahara, T. Matsui, and S. Nakamura, "Out-of-domain detection based on confidence measures from multiple topic classification," in *Proc. ICASSP*, 2004, vol. 1, pp. 757–760.

[32] A. Lee, T. Kawahara, and K. Shikano, "Julius- an open source real-time large vocabulary recognition engine," in *Proc. Eurospeech*, 2001, pp. 1691–1694.

[33] B. Carlson, "Unsupervised topic clustering of switchboard speech messages," in *Proc. ICASSP*, 1996, pp. 315–318.

**Ian Lane** (M'97) received the B.Tech. degree in information engineering from Massey University, New Zealand, in 2000. He was awarded a Monbukagakusho scholarship to study in the School of Informatics, Kyoto University, Kyoto, Japan, where he received the M.E. degree in 2003 and is currently pursuing the Ph.D. degree.

He is an Intern Researcher at ATR Spoken Language Communication Research Laboratories, Kyoto.

**Tatsuya Kawahara** (M'91) received the B.E., M.E., and Ph.D. degrees in information science, from Kyoto University, Kyoto, Japan, in 1987, 1989, and 1995, respectively.

In 1990, he became a Research Associate with the Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ. Currently, he is a Professor at the Academic Center for Computing and Media Studies, Kyoto University. He is also an Invited Researcher at ATR Spoken Language Communication Research Laboratories. He has published more than 100 technical papers covering speech recognition, confidence measures, and spoken dialogue systems. He has been managing several speech-related projects in Japan including a free large vocabulary continuous speech recognition software project (http://www.julius.sourceforge.jp/).

Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. Since 2003, he has been a member of the IEEE SPS Speech Technical Committee.

**Tomoko Matsui** (M'91) received the Ph.D. degree from the Computer Science Department, Tokyo Institute of Technology, Tokyo, Japan, in 1997.

From 1988 to 2002, she was with NTT, where she worked on speaker and speech recognition. From 1998 to 2002, she was with the Spoken Language Translation Research Laboratory, ATR, Kyoto, Japan, as a Senior Researcher and worked on speech recognition. From January to June 2001, she was an Invited Researcher in the Acoustic and Speech Research Department, Bell Laboratories, Murray Hill, NJ, working on finding effective confidence measures for verifying speech recognition results. She is presently an Associate Professor in the Institute of Statistical Mathematics, Tokyo, working on statistical modeling for speech and speaker recognition applications.

Dr. Matsui received the paper award of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) in 1993.

**Satoshi Nakamura** (M'89) received the B.S. degree in electronic engineering from Kyoto Institute of Technology, Kyoto, Japan, in 1981 and the Ph.D. degree in information science from Kyoto University in 1992.

From 1981 to 1993, he was with the Central Research Laboratory, Sharp Corporation, Nara, Japan. From 1986 to 1989, he worked with ATR Interpreting Telephony Research Laboratories. From 1994 to 2000, he was an Associate Professor of the Graduate School of Information Science at Nara Institute of Science and Technology, Japan. In 1996, he was a Visiting Research Professor of the CAIP Center, Rutgers University, Newark, NJ. He is currently the Director at ATR Spoken Language Communication Laboratories, Japan. He has been an Honorary Professor at the University of Karlsruhe, Karlsruhe, Germany, since 2004. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays. He has published more than 100 papers in acoustics, speech and natural languages processing. He served as an editor of the *Journal of IEICE Information* from 2000 to 2002.

Dr. Nakamura received the Awaya Award from the Acoustical Society of Japan in 1992, and the Interaction 2001 Best Paper Award from the Information Processing Society of Japan in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society from 2001 to 2004.