Topic Classification and Verification Modeling for Out-of-Domain Utterance Detection

Ian R. Lane[1,2], *Tatsuya Kawahara*[1,2], *Tomoko Matsui*[3,2], *Satoshi Nakamura*[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
[3]The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Mitato-ku, Tokyo 106-8569, Japan

Abstract

The detection and handling of OOD (out-of-domain) user utterances are significant problems for spoken language systems. We approach these problems by applying an OOD detection framework, combining topic classification and in-domain verification. In this paper, we compare the performance of three topic classification modeling schemes: *1-vs-all*, where a single classifier is trained for each topic; weighted *1-vs-all*; and *1-vs-1*, which combines multiple pair-wise classifiers. We also compare the performance of a linear discriminate verifier and nonlinear SVM-based verification. In an OOD detection task as a front-end for speech-to-speech translation, detection performance was comparable for all classification schemes, indicating that the simplest *1-vs-all* approach is sufficient for this task. SVM-based in-domain verification was found to provide a significant reduction in detection errors compared to a linear discriminate model. However, when the training and testing scenarios differ, the SVM approach was not robust, while the linear discriminate model remained effective.

1. Introduction

Spoken language systems are designed to operate over limited application domains for improved recognition performance. The domain is typically defined by the back-end system. Users, especially novice users, however, often do not have an exact concept of the application domain and may attempt utterances that cannot be handled by the system. These are referred to as OOD (out-of-domain) utterances in this paper. The definition of out-of-domain is dependent on the type of spoken language system. Definitions for three typical systems are described in Table 1.

The detection of OOD utterances is vital to improve system usability. OOD detection will allow effective user feedback to be generated enabling users to determine whether they should re-attempt the current task after it is confirmed as in-domain, or to halt the current task after being informed that it cannot be handled by the system. For example, a speech-to-speech translation system with OOD detection will interact with a user as shown in Figure 1. In the first example (A), an in-domain utterance could not be accurately translated by the back-end system; in this case the user is requested to re-phrase the input utterance, making translation possible. However, if an OOD utterance is encountered as in example B, it cannot be handled correctly regardless of how it is re-phrased. In such cases the user should be informed that the utterance is OOD and provided with a detailed explanation of the application domain.

Table 1: Definition of out-of-domain for various systems

System	Out-of-Domain definition
Spoken Dialogue	User's query does not relate to the 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 offered topic

Example A: In-domain utterance, re-phrased

User	Excuse me, uh, could you tell me where I can find a taxi stand ?
Sys	"Please re-phrase that." *Utterance detected as in-domain, translation confidence low*
User	I'm trying to find a taxi, a taxicab downtown. *Utterance in-domain, translation confidence high*

Example B: Out-of-domain utterance encountered

User	I can't get my computer to work. *Utterance detected as out-of-domain*
Sys	"I'm sorry, only travel related topics can be handled."

Figure 1: OOD detection for speech-to-speech translation

In previous work [1], we studied a framework for detecting OOD utterances which does not require OOD data during training. In this approach, the domain is assumed to consist of multiple sub-domain topics, such as call destinations in call-routing, sub-topics in translation systems, and sub-domains in complex dialogue systems. OOD detection is performed by first calculating classification scores for all topic classes and then applying an in-domain verification model to this vector, resulting in an OOD decision. In [1], a linear discriminate model-based verifier was trained using deleted interpolation on the in-domain data, providing acceptable detection accuracy even when no OOD training data is available.

In this paper, we compare the system performance for various topic classification schemes and in-domain verification models. We also investigate the portability of the system by evaluating OOD detection accuracy when training and testing scenarios differ.

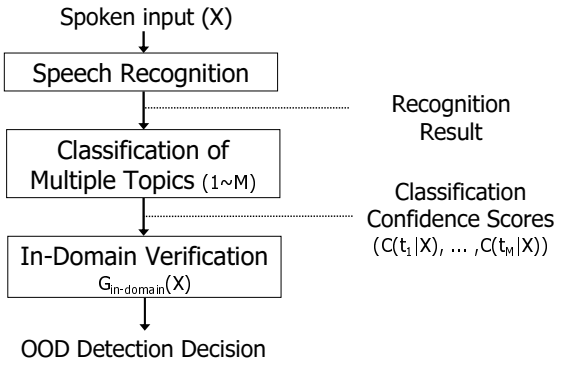

Figure 2: OOD detection based on topic classification

2. System Overview

In the OOD detection 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-labeled appropriately. These data are then used to train topic classification models.

An overview of the system is shown in Figure 2. First, speech recognition is performed by applying a generalized language model that covers all topics, and a recognition hypothesis is generated. Next, topic classification models are applied to the recognition hypothesis to create a vector of their confidence scores $\big(C(t_1|X), \ldots, C(t_M|X)\big)$. Finally, an in-domain verification model $G_{in-domain}(X)$ is applied to this vector resulting in a binary OOD decision. The performance of this framework is affected by two aspects: the accuracy of topic classification and the discriminative ability of the in-domain verification model. These two aspects are described in detail in the following sections.

3. Topic Classification

Topic classification involves creating a vector of topic classification confidence scores for each input utterance. Each component within the vector corresponds to the confidence score for an individual topic class. Classification is based on a vector space model approach, similar to the method described in [2]. First, sentences are projected to large-dimension vectors (consisting of 10,000 - 70,000 features), where each vector component corresponds to the occurrence of a specific word or N-gram feature as described in Section 3.1. Within this vector space, one or more SVM (support vector machine) classifiers are then trained to model each topic class.

In this paper, three topic classification schemes are evaluated: *1-vs-all*, *1-vs-all-weighted*, and *1-vs-1*. These are described in Sections 3.2 - 3.4 below.

3.1. Topic Classification Features

Topic classification accuracy is affected not only by the classification scheme but also by the classification feature set used. We investigate the system performance for various feature sets. Features consisting of word baseform (word token with no tense information; all variants are merged), full-word (surface form of words, including variants), and word+POS (part-of-speech) tokens are compared. The inclusion of N-gram features that combine multiple neighboring tokens is also investigated. Appropriate cutoffs are applied during training to remove features with low occurrence.

3.2. 1-vs-all Topic Classification

The *1-vs-all* scheme involves creating a single SVM classifier for each topic class. In this approach, a linear SVM classifier (SVM_i) is trained to discriminate between the current topic class (i) and all other classes. During training, sentences that occur in the training set of that class are used as positive examples and the remainder of the training set is used as negative examples. Class size is not considered during training.

The classification confidence score of a topic class is calculated as shown in Equation 1. First, the perpendicular distance from the SVM hyperplane SVM_i to the input utterance X ($dist_{SVM_i}(X)$) is calculated. This is the minimum distance to the class boundary as defined by the SVM model. Distance is positive if the input utterance is in-class or negative otherwise. A classification confidence score $C(t_i|X)$, in the range $[0, 1]$, is then calculated by applying a sigmoid function to this distance.

$$C(t_i|X) = \ sigmoid\big(dist_{SVM_i}(X)\big) \tag{1}$$

$sigmoid()$: Sigmoid function
$dist_{SVM_i}(X)$: Perpendicular distance from SVM hyperplane (SVM_i) to input utterance X

3.3. 1-vs-all-weighted Topic Classification

For the *1-vs-all-weighted* scheme, training is identical to the *1-vs-all* case, except that a misclassification penalty is applied during training to compensate for unbalanced class sizes. In this approach, the misclassification penalties for small topic classes are increased, resulting in similar misclassification rates for all topics. The classification confidence score is calculated as for the *1-vs-all* case (Equation 1).

3.4. 1-vs-1 Topic Classification

Both of the above schemes train a single SVM classifier for each topic class. Classification accuracy may be improved by applying a more complex classification model consisting of multiple SVM classifiers. In the *1-vs-1* approach, each topic class is modeled with a piece-wise linear boundary composed of multiple SVM classifiers. Each topic class is modeled by a set of $(M-1)$ pair-wise classifiers $(SVM_{i,1}, \ldots, SVM_{i,j}, \ldots, SVM_{i,M})$, where $i \neq j$, M is the total number of topic classes. Each classifier ($SVM_{i,j}$) is trained to discriminate between a pair of topic classes: the current topic (i) and one other topic class (j).

The classification confidence score is calculated as shown in Equation 2. The distance is calculated for each SVM classifier and the minimum is output.

$$C(t_i|X) = sigmoid\Big(\min_j \big(dist_{SVM_{i,j}}(X)\big)\Big) \tag{2}$$

4. In-domain Verification

In the final stage of the framework, an OOD decision is generated by applying an in-domain verification model $G_{in-domain}(X)$ to the vector of classification confidence scores. We compare the performance of two verification mod-

els: a linear discriminate model trained using minimum classification error learning, and a nonlinear SVM-based model.

4.1. Linear Discriminate Model based Verification

In [1], we investigated a linear discriminant model trained using deleted interpolation on the in-domain data. In this paper, we investigate the system performance and robustness when both in-domain and OOD training data are available.

The linear discriminate model (Equation 3) involves applying linear discriminant weights $(\lambda_1, \ldots, \lambda_M)$ to the vector of confidence scores $\big(C(t_1|X), \ldots, C(t_M|X)\big)$ generated during topic classification. A threshold (φ) is applied to the resulting value to obtain a binary decision of in-domain or OOD.

$$G_{\text{in-domain}}(X) = \begin{cases} 1 & \text{if} \sum_{i=1}^{M} \lambda_i C(t_i|X) \geq \varphi \\ & \text{(in-domain)} \\ 0 & \text{otherwise.} \\ & \text{(OOD)} \end{cases} \quad (3)$$

$C(t_i|X)$: Classification score of topic i for input utterance X
M: Total number of topic classes

The discriminant weights $(\lambda_1, \ldots, \lambda_M)$ are estimated using the GPD (gradient probabilistic descent) algorithm as described in [4]. During training, in-domain data are used as positive training examples, and OOD data as negative examples.

4.2. Non-linear SVM-based Verification

For comparison, we introduce a nonlinear SVM-based verifier. An RBF (radial basis function) kernel as shown in Equation 4 is applied, enabling a complex nonlinear verification model to be generated. The RBF variable gamma (γ) is automatically selected during training using the method described in [3]. An OOD decision is obtained by applying a threshold to the resulting distance from the SVM verifier.

$$K_{RBF}(x_i, x_j) = exp(-\gamma \| x_i - x_j \|^2) \quad (4)$$

5. Experimental Evaluation

The performance of the OOD detection framework was evaluated for a speech-to-speech translation front-end. Evaluation was performed using the ATR BTEC corpus [5]. This corpus covers a travel domain and consists of 12 topic classes. In this evaluation, one topic, "*shopping*", is set as OOD of the system. The remaining 11 topics, consisting of 149,540 sentences are in-domain, and these data are used to train both the language model for speech recognition and the topic classification models. Test sets of 1,852 in-domain and 138 OOD utterances are used for evaluation.

OOD detection performance was evaluated by performing five-fold cross validation on the test set. In this method, 80% of the data was used to train the in-domain verification model and the remaining 20% was used for evaluation. This process was repeated five times and the average EER (equal error rate) was calculated. EER is determined by setting the threshold of the verification model so that the false rejection of in-domain data and false acceptance of OOD data are equal.

Table 2: Comparison of feature sets

ID	Feature set	no. features	OOD detection EER(%)	
			LDM	SVM
A	base-form	8771	25.2%	17.8%
B	word+POS	9899	23.7%	18.1%
C	word	10006	23.3%	15.6%
D	word,2-gram	40754	20.6%	16.1%
E	word,2,3-gram	73065	19.1%	15.3%

LDM: Linear discriminate model
SVM: SVM-based verification

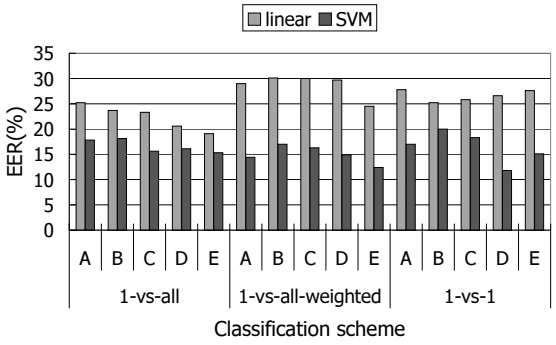

Figure 3: OOD detection accuracy for transcriptions

5.1. Comparison of Verification Models

First, the system performance was evaluated for combinations of feature sets and verification models using the manual transcriptions of the test sets. The OOD detection EER of each case is shown in Table 2.

The SVM verifier significantly outperformed the linear model for all feature sets. This indicates that the SVM-based model is more effective for OOD detection when a suitable set of OOD training data is available.

Context-based (2-gram, 3-gram) features improved OOD detection accuracy for the linear model case, however, there was no significant improvement when combined with SVM-based verification. It is assumed that the detection accuracy of the SVM verifier is close to optimal even with a limited set of topic classification features, thus, the addition of context-based features do not improve system performance.

5.2. Investigation of Topic Classification Models

Next, the three topic classification schemes described in Section 3 were evaluated. Classification models were trained for each scheme by applying the feature sets investigated in Section 5.1. The OOD detection performance of each scheme, combined with both linear and SVM-based verification models was evaluated (Figure 3).

When a linear verification model is applied, classification-based on the *1-vs-all* scheme incorporating word, 2-gram and 3-gram features provides the best performance with an OOD detection EER of 19.1%. Applying the *1-vs-all-weighted* scheme reduces performance compared to the non-weighted approach. As the distribution of class sizes is similar for both training and testing, there is no need for compensation.

For both SVM and linear discriminate verification, the *1-vs-1* topic classification scheme provides no significant reduction

Table 3: Speech recognition performance

	# Utt.	WER(%)	SER(%)	OOV(%)
In-domain	1852	7.26	22.4	0.71
Out-of-domain	138	12.49	45.3	2.56

WER: Word error rate SER: Sentence error rate
OOV: Out of vocabulary rate

Figure 4: OOD detection accuracy for open data

in OOD detection errors compared to the *1-vs-all* case. The number of dimensions of the feature space is extremely large (10,000 - 70,000 dimensions), thus the simple *1-vs-all* classification scheme is sufficient.

Compared to the linear discriminate case, SVM-based verification improves OOD detection accuracy for all combinations of feature sets and topic classification schemes. Detection performance is comparable for all topic classification schemes when SVM-based verification is used, indicating the effectiveness of SVM for a wide range of input features.

5.3. ASR-based OOD detection

Next, the OOD detection performance is evaluated on automatic speech recognition results. Speech recognition is performed with the Julius recognition engine [6]. The recognition performance for the in-domain (ID) and OOD sets are shown in Table 3. *1-vs-all* topic classification is applied based on the feature set incorporating word, 2-gram and 3-gram features. The system performance is evaluated for both linear and SVM-based verification. OOD detection EER of 22.6% and 16.1% are gained respectively for the two models. Compared to the OOD detection performance on the manual transcriptions (19.1% and 15.3% respectively) there is only a small degradation in accuracy, suggesting that the OOD detection framework is robust against speech recognition errors.

5.4. Evaluation of System Portability and Robustness

Finally, the portability of the system is evaluated by testing the system on open data from a topic that was not available during training. A *business* test set is introduced which consists of utterances that occur in business travel situations. The test set consists of 880 in-domain utterances (related to the travel domain) and 682 OOD utterances. *1-vs-all* topic classification was applied and OOD detection performance was compared for linear and SVM-based verification. For reference, five-fold cross validation was applied to the *business* test set, as described in Section 5.3. The system was then trained on the *shopping* test set and the *business* test set was used for open evaluation. The system performance when training and testing on the same corpus (cross validation) and when training on one corpus (*shopping*) and evaluating on the other (*business*) is shown in Figure 4.

Although the SVM verification model offers higher detection accuracy with the cross validation scheme when trained and evaluated on the same topic, its performance is very poor when the evaluation data are open. For adequate performance the SVM approach requires sufficient OOD training data. The linear discriminate approach, on the other hand, realizes improved robustness and portability because it adopts a much simpler verification model.

6. Conclusions

We evaluated topic classification models and in-domain verifiers for the proposed OOD detection framework. In this framework confidence scores of multiple topic classification are calculated and a in-domain verifier is applied resulting in an OOD decision. Three topic classification modeling schemes based on SVM (support vector machine) classifiers were compared. OOD detection performance was comparable for all classification schemes, indicating that the simplest 1-vs-all approach is sufficient for this task. For in-domain verification modeling, we compared a linear discriminate model and a non-linear SVM-based model. Although the SVM-based method provided better OOD detection when sufficient training data were available, it significantly degraded the performance for open data, while the linear discriminate model realized robust performance.

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] 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.

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

[3] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A practical guide to support vector classification. July, 2003. http://www.csie.ntu.edu.tw/ cjlin/libsvm/.

[4] S. Katagiri, C.-H. Lee, and B.-H. Juang, New discriminative training algorithm based on the generalized probabilistic descent method. In Proc. *IEEE workshop NNSP, pp. 299-300*, 1991.

[5] 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.

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