OUT-OF-DOMAIN DETECTION BASED ON CONFIDENCE MEASURES FROM MULTIPLE TOPIC CLASSIFICATION

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

One significant problem for spoken language systems is how to cope with users' OOD (out-of-domain) utterances which cannot be handled by the back-end system. In this paper, we propose a novel OOD detection framework, which makes use of classification confidence scores of multiple topics and trains a linear discriminant in-domain verifier using GPD. Training is based on deleted interpolation of the in-domain data, and thus does not require actual OOD data, providing high portability. Three topic classification schemes of word N-gram models, LSA, and SVM are evaluated, and SVM is shown to have the greatest discriminative ability. In an OOD detection task, the proposed approach achieves an absolute reduction in EER of 6.5% compared to a baseline method based on a simple combination of multiple-topic classifications. Furthermore, comparison with a system trained using OOD data demonstrates that the proposed training scheme realizes comparable performance while requiring no knowledge of the OOD data set.

1. INTRODUCTION

Most spoken language systems, excluding general-purpose dictation systems, operate over definite domains as a user interface to a service provided by the back-end system. However, users, especially novice users, do not always have an exact concept of the domains served by the system. Thus, they often attempt utterances that cannot be handled by the system. These are referred to as OOD (out-of-domain) in this paper. Definitions of OOD for three typical spoken language systems are described in Table 1.

For an improved interface, spoken language systems should predict and detect such OOD utterances. In order to predict OOD utterances, the language model should allow some margin in its coverage. A mechanism is also required for the detection of OOD utterances, which is addressed in this paper. Performing OOD detection will improve the system interface by enabling users to determine whether to reattempt the current task after being confirmed as in-domain, or to halt attempts due to being OOD. For example, in a speech-to-speech translation system, an utterance may be in-domain but unable to be accurately translated by the back-end system; in this case the user is requested to re-phrase the input utterance, making translation possible. In the case of an OOD utterance, however, re-phrasing will not improve translation, so the user should be informed that the utterance is OOD and provided with a list of tractable domains.

Table 1. Definitions of Out-of-domain 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 offered topic

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 successfully applied is call routing tasks such as that described in [3]. In this work, classification models are trained for each call destination, and a garbage model is explicitly trained to detect OOD utterances. To 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 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 the proposed 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 confidence scores for all in-domain topic classes and then applying an in-domain verification model to this confidence vector, which results in an OOD decision. The verification model is trained using GPD (gradient probabilistic descent) and deleted interpolation, allowing the system to be developed by using only in-domain data.

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-labeled appropri-

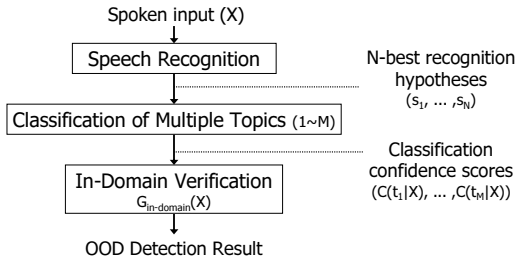

Fig. 1. Topic Classification based OOD Detection

ately. These data are then used to train topic classification models. Topic classification is also useful for improving ASR performance by applying topic-dependent language models. We demonstrated the effectiveness of such an approach in [4].

An overview of the OOD detection framework is shown in Figure 1. First, speech recognition is performed by applying a generalized language model that covers all in-domain topics, 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 for each topic class based on these hypotheses. Finally, OOD detection is performed by applying an in-domain verification model $G_{in-domain}(X)$ to the resulting confidence vector. The overall performance of the proposed approach is affected by the accuracy of the topic classification method and the in-domain verification model. These aspects are described in detail in the following sections.

3. TOPIC CLASSIFICATION

In this paper three topic classification schemes are evaluated: topic-dependent word N-gram, LSA (latent semantic analysis), and SVM (support vector machines). Based on a given feature set, topic models are trained using the above methods. Topic classification is performed and confidence scores (in the range $[0, 1]$) are calculated by applying a sigmoid transform to these results. When classification is applied to an N-best speech recognition result, confidence scores are calculated as shown in Equation 1. Topic classification is applied independently to each N-best hypothesis, and these are linearly combined by weighting each with the posterior probability of that hypothesis given by ASR.

$$C(t_j|X) = \sum_{i=1}^{N} p(s_i|X) C(t_j|s_i) \tag{1}$$

$C(t_j|X)$: confidence score of topic t_j for input utterance X
$p(s_i|X)$: posterior probability of i-th best sentence hypothesis s_i by ASR
N: number of N-best hypotheses

3.1. Topic Classification Features

Various feature sets for topic classification are investigated. A feature vector consists of 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) tokens. 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. Topic-dependent Word N-gram

In this approach, N-gram language models are trained for each topic class. Classification is performed by calculating the log-likelihood of each topic model for the input sentence. Topic classification confidence scores are calculated by applying a sigmoid transform to this log-likelihood measure.

3.3. Latent Semantic Analysis

LSA (latent semantic analysis) [5] is a popular technique for topic classification. Based on a vector space model, each sentence is represented as a point in a large dimension space, where vector components relate to the features described in Section 3.1. Because the vector space tends to be extremely large (10,000-70,000 features), traditional distance measures such as the cosine distance become unreliable. To improve performance, SVD (singular value decomposition) is applied to reduce the large space to 100-300 dimensions. Each topic class is represented as a single document vector composed of all training sentences, and projected to this reduced space.

Classification is performed by projecting the vector representation of the input sentence to the reduced space and calculating the cosine distance between this vector and each topic class vector. The resulting distance is normalized by applying a sigmoid transform generating classification confidence scores.

3.4. Support Vector Machines

SVM (support vector machines) [6] is another popular classification technique. Using a vector space model, SVM classifiers are trained for each in-domain topic class. Sentences that occur in the training set of that topic are used as positive examples and the remainder of the training set is used as negative examples.

Classification is performed by feeding the vector representation of the input sentence to each SVM classifier. The perpendicular distance between this vector and each SVM hyperplane is used as the classification measure. This value is positive if the input sentence is in-class and negative otherwise. Again, confidence scores are generated by applying a sigmoid transform to this distance.

4. IN-DOMAIN VERIFICATION

The final stage of OOD detection consists of applying an in-domain verification model $G_{in-domain}(X)$ to the vector of confidence scores generated during topic classification. We adopt a linear discriminant model (Eqn. 2). Linear discriminant weights $(\lambda_1, \ldots, \lambda_M)$ are applied to the confidence scores from topic classification $(C(t_1|X), \ldots, C(t_M|X))$, and a threshold (φ) is applied to obtain a binary decision of in-domain or OOD.

$$G_{in-domain}(X) = \begin{cases} 1 \text{ if} \sum_{j=1}^{M} \lambda_j C(t_j|X) \geq \varphi & \text{(in-domain)} \\ 0 \text{ otherwise.} & \text{(OOD)} \end{cases} \tag{2}$$

$C(t_j|X)$: confidence score of topic t_j for input utterance X
M: number of topic classes

4.1. Training using Deleted Interpolation

The in-domain verification model is trained using only in-domain data. An overview of the proposed training method combining GPD (gradient probabilistic descent) [7] and deleted interpolation

Table 2. 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 iterations

Table 3. Experiment Corpus

Domain: Basic Travel Expressions
  In-Domain: 11 topics (*transit*, *accommodation*, ...)
  OOD: 1 topic (*shopping*)
Training Set: 11 topics, 149540 sentences (in-domain data only)
Lexicon Size: 17000 words
Test set: **In-Domain**: 1852 utterances
  OOD: 138 utterances

is given in Table 2. Each topic is iteratively set to be temporarily OOD, and the classifier corresponding to this topic is removed from the model. The discriminant weights of the remaining topic classifiers are estimated using GPD. In this step, the temporary OOD data is used as negative training examples, and a balanced set of the remaining topic classes are used as positive (in-domain) examples. Upon completion of estimation by GPD, the final model weights are calculated by averaging over all interpolation steps. In the experimental evaluation, a topic-independent class "*basic*" covering general utterances exists, which is not removed during deleted interpolation.

4.2. Incorporation of Topic-dependent Verifier

Improved OOD detection accuracy can be achieved by applying more elaborate verification models. In this paper, a model consisting of multiple linear discriminant functions is investigated. Topic dependent functions are added for topics not modeled sufficiently. Their weights are trained specifically for verifying that topic. For verification, the topic with maximum classification confidence is selected, and a topic-dependent function is applied if one exists, otherwise a topic-independent function (Eqn. 2) is applied.

5. EXPERIMENTAL EVALUATION

The ATR BTEC corpus [8] is used to investigate the performance of the proposed approach. An overview of the corpus is given in Table 3. In this experiment, we use "*shopping*" as OOD of the speech-to-speech translation system. The training set consisting of 11 in-domain topics is used to train both the language model for speech recognition and the topic classification models. Recognition is performed with the Julius recognition engine.

The recognition performance for the in-domain (ID) and OOD test sets are shown in Table 4. Although the OOD test set has much greater error rates and out-of-vocabulary rate compared with the in-domain test set, more than half of the utterances are correctly recognized, since the language model covers the general travel domain. This indicates that the OOD set is related to the in-domain task, and discrimination between these sets will be difficult.

System performance is evaluated by the following measures:

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

Table 4. 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

Table 5. Comparison of Feature Sets & Classification Models

Method	Token Set	Feature Set	# Feat.	EER(%)
SVM	base-form	1-gram	8771	29.7
SVM	full-word	1-gram	9899	23.9
SVM	word+POS	1-gram	10006	23.3
SVM	word+POS	1,2-gram	40754	21.7
SVM	word+POS	1,2,3-gram	73065	19.6
LSA	word+POS	1-gram	10006	23.3
LSA	word+POS	1,2-gram	40754	24.1
LSA	word+POS	1,2,3-gram	73065	23.0
NGRAM	word+POS	1-gram	10006	24.8
NGRAM	word+POS	1,2-gram	40754	25.2
NGRAM	word+POS	1,2,3-gram	73065	24.2

SVM:Support Vector Machines LSA:Latent Semantic Analysis
NGRAM:Topic-dependent Word N-gram

5.1. Evaluation of Topic Classification and Feature Sets

First, the discriminative ability of various feature sets as described in Section 3.1 were investigated. Initially, SVM topic classification models were trained for each feature set. A closed evaluation was performed for this preliminary experiment. Topic classification confidence scores were calculated for the in-domain and OOD test sets using the above SVM models, and used to train the in-domain verification model using GPD. During training, in-domain data were used as positive training examples, and OOD data were used as negative examples. Model performance was evaluated by applying this closed model to the same confidence vectors used for training. The performance in terms of EER is shown in the first section of Table 5.

The EER when word-baseform features were used was 29.7%. Full-word or word+POS features improved detection accuracy significantly: with EERs of 23.9% and 23.3%, respectively. The inclusion of context-based 2-gram and 3-gram features further improved detection performance. A minimum EER of 19.6% was obtained when 3-gram features were incorporated.

Next, LSA and N-gram-based classification models were evaluated. Both approaches showed lower performance than SVM, and the inclusion of context-based features did not improve performance. SVM with a feature set containing 1-, 2-, and 3-gram offered the lowest OOD detection error rate, so it is used in the following experiments.

5.2. Deleted Interpolation-based Training

Next, performance of the proposed training method combining GPD and deleted interpolation was evaluated. We compared the OOD detection performances of the proposed method (proposed), a reference method in which the in-domain verification model was trained using both in-domain and OOD data (as described in Section 5.1) (closed-model), and a baseline system. In the baseline system, topic detection was applied and an utterance was classified as OOD if all binary SVM decisions were negative. Other-

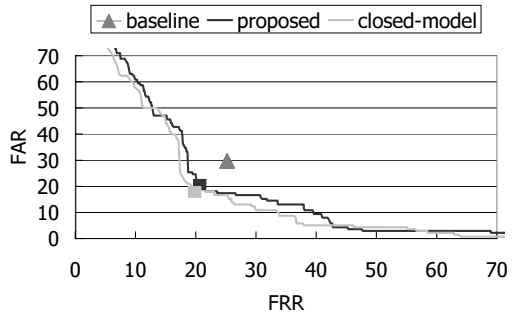

Fig. 2. OOD Detection Performance on Correct Transcriptions

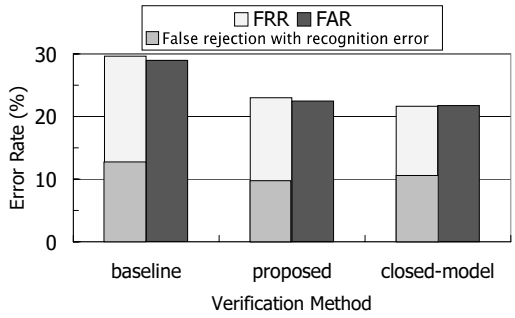

Fig. 3. OOD Detection Performance on ASR Result

wise it was classified as in-domain. The ROC graph of the three systems obtained by altering the verification threshold (φ in Eqn. 2) is shown in Figure 2.

The baseline system has a FRR of 25.2%, a FAR of 29.7%, and an EER of 27.7%. The proposed method provides an absolute reduction in EER of 6.5% compared to the baseline system. Furthermore, it offers comparable performance to the closed evaluation case (21.2% vs. 19.6%) while being trained with only in-domain data. This shows that the deleted interpolation approach is successful in training the OOD detection model in the absence of OOD data.

5.3. Evaluation with ASR Results

Next, the performances of the above three systems were evaluated on a test set of 1990 spoken utterances. Speech recognition was performed and the 10-best recognition results were used to generate a topic classification vector. The FRR, FAR and percentage of falsely rejected utterances with recognition errors are shown in Figure 3.

The EER of the proposed system when applied to the ASR results is 22.7%, an absolute increase of 1.5% compared to the case for the correct transcriptions. This small increase in EER suggests that the system is strongly robust against recognition errors. Further investigation showed that the falsely rejected set had a SER of around 43%, twice that of the in-domain test set. This suggests that utterances that incur recognition errors are more likely to be rejected than correctly recognized utterances.

5.4. Effect of Topic-dependent Verification Model

Finally, the topic-dependent in-domain verification model described in Section 4.2 was also incorporated. Evaluation was performed on spoken utterances as in the above section. The addition of a topic-dependent function (for the topic "*basic*") reduced the EER to 21.2%. The addition of further topic-dependent functions, however, did not provide significant improvement in performance over the two function case. The topic class "*basic*" is the most vague and is poorly modeled by the topic-independent model. A topic-dependent function effectively models the complexities of this class.

6. CONCLUSIONS

We proposed a novel OOD (out-of-domain) detection method based on confidence measures from multiple topic classification. A novel training method combining GPD and deleted interpolation was introduced to allow the system to be trained using only in-domain data. Three classification methods were evaluated (topic dependent word N-gram, LSA and SVM), and SVM-based topic classification using word and N-gram features proved to have the greatest discriminative ability.

The proposed approach reduced OOD detection errors by 6.5% compared to the baseline system based on a simple combination of binary topic classifications. Furthermore, it provides similar performance to the same system trained on both in-domain and OOD data (EERs of 21.2% and 19.6%, respectively) while requiring no knowledge of the OOD data set. Addition of a topic dependent verification model provides a further reduction in detection errors.

Acknowledgements: The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus".

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

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

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

[5] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. In *Journ. of the American Society for information science, 41, pp. 391-407*, 1990.

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

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

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