

# Dialogue Speech Recognition by Combining Hierarchical Topic Classification and Language Model Switching

Ian R. LANE<sup>†,††</sup>, Tatsuya KAWAHARA<sup>†,††</sup>, Tomoko MATSUI<sup>††,†††</sup>,  
and Satoshi NAKAMURA<sup>††</sup>, *Members*

**SUMMARY** An efficient, scalable speech recognition architecture combining topic detection and topic-dependent language modeling is proposed for multi-domain spoken language systems. In the proposed approach, the inferred topic is automatically detected from the user's utterance, and speech recognition is then performed by applying an appropriate topic-dependent language model. This approach enables users to freely switch between domains while maintaining high recognition accuracy. As topic detection is performed on a single utterance, detection errors may occur and propagate through the system. To improve robustness, a hierarchical back-off mechanism is introduced where detailed topic models are applied when topic detection is confident and wider models that cover multiple topics are applied in cases of uncertainty. The performance of the proposed architecture is evaluated when combined with two topic detection methods: unigram likelihood and SVMs (Support Vector Machines). On the ATR Basic Travel Expression Corpus, both methods provide a significant reduction in WER (9.7% and 10.3%, respectively) compared to a single language model system. Furthermore, recognition accuracy is comparable to performing decoding with all topic-dependent models in parallel, while the required computational cost is much reduced.

**key words:** *Speech Recognition, Topic Detection, Topic-Dependent Language Modeling, Support Vector Machines, Multi-domain Spoken Dialogue*

## 1. Introduction

Speech is a natural communication medium, and is thus an efficient interface for human-machine communications. In recent years, there has been significant growth in the development and commercial deployment of interactive spoken language systems. Applications include spoken dialogue systems for guidance and transactions [1]–[5], and more recently speech-to-speech translation systems [6]–[8]. To enhance their performance, these systems are specifically designed to operate over limited and definite domains. By limiting operation to a single application task, robust speech recognition can be realized. This approach has led to the development of a large number of spoken dialogue systems. For example, the MIT GALAXY [9] framework

has been used to develop the JUPITER [1], DINEX [2] and VOYAGER [3] spoken dialogue systems, which provide access to weather, restaurant and urban navigation information, respectively.

Limiting operation to a single task domain, however, forces users to make use of several independent systems when they require information from multiple domains (for example, transit information to a particular city, as well as the weather forecast for that city.) For improved usability, systems should operate across multiple domains allowing users to gain the required information quickly and efficiently within a single interaction.

Commercial “Voice Portal” systems provide the simplest spoken language interface for information retrieval over multiple domains. In these systems, a pre-defined set of “command keywords” are used to traverse an information hierarchy to obtain the required information. This approach only requires the recognition of a limited set of keywords and thus realizes reasonable performance. However, the usability of such systems is limited as users require knowledge of both the system's keywords and information structure before they can effectively use the system. Expert users are not favored, either, as they are forced to traverse the information hierarchy, even when they have full knowledge of the system.

An alternative approach to multi-domain spoken dialogue includes systems developed for the DARPA “Communicator” project [10], [11]. These systems cover the travel domain allowing users to retrieve up-to-date flight schedules, flight pricing, hotel information, and rental car availability. They improve over the “Voice Portal” approach by applying a conversational front-end rather than just keyword recognition. These systems typically adopt a single recognition front-end that provides coverage over all of the sub-domains contained within the system. A similar approach was also applied in [12] for an office assistant spoken dialogue system that operates over a large number of sub-domains.

When performing speech recognition over multiple domains, topic- or sub-task-dependent language modeling increases both the accuracy and efficiency of the system. However, current dialogue systems that use multiple topic-dependent language models typically adopt

<sup>†</sup>The authors are with the School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

<sup>††</sup>The authors are with the Department of Acoustic and Speech Research, Advance Telecommunications Research Institute International, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

<sup>†††</sup>The author is with the Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Mitato-ku, Tokyo 106-8569, Japan

a system initiative approach [13], [14] where the appropriate LM is applied based on the system’s prompt, determined by the dialogue flow of the system. Increased usability can be achieved by allowing users to switch between domains as in [15], but in this case, users must explicitly state the domain they require before making a query. A multi-domain spoken dialogue system based on this approach would behave as follows. (S stands for system and U for User)

- S) Welcome to the Kyoto City Information Portal.  
 What system do you require; tourist, restaurant or bus?  
 U) Tourist information, please.  
 S) You are now in the Kyoto Tourist Information System.  
 U) What time is the Golden Pavilion open until?  
 S) The Golden Pavilion is open from 8:30am until 5:00pm, every day of the week.  
 U) Restaurant information.  
 S) You are now in the Kyoto Restaurant Information System.  
 U) Japanese style restaurants near the Golden Pavilion?  
 S) There are 2 Japanese restaurants in that vicinity, ...

In this approach, the number of dialogue turns is unreasonably large due to the overhead required to switch between domains. Rather than explicitly stating which domain is required, the system should automatically detect it from the users’ utterance. For this purpose, approaches used in call routing can be applied [16], [17]. In this study, we propose a recognition architecture combining topic detection and topic-dependent language modeling. The inferred domain is automatically detected from the user’s utterance, and speech recognition is then performed with an appropriate topic-dependent language model. This allows the user to seamlessly switch between domains while maintaining high recognition accuracy. Based on this approach, the above spoken dialogue system would behave as follows.

- S) Welcome to the Kyoto City Information Portal.  
 Please query the system on tourist, restaurant or bus information.  
 U) What time is the Golden Pavilion open until?  
 S) The Golden Pavilion is open from 8:30am until 5:00pm, every day of the week.  
 U) Are there any good Japanese restaurants near there?  
 S) There are 2 Japanese restaurants in that vicinity, ...

This approach significantly reduces the number of dialogue turns required. As the domain of each user utterance is automatically detected, the overhead required to explicitly switch between domains is eliminated. Recognition accuracy is also maintained as topic-dependent recognition is applied. An alternative approach for topic dependent recognition is to perform decoding with multiple topic-dependent language models in parallel, however this approach requires large

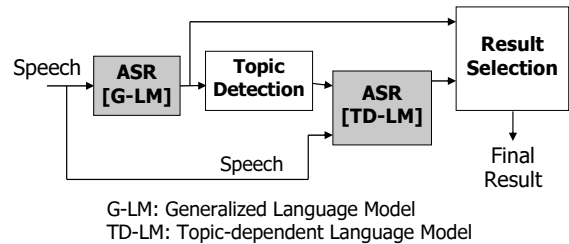


Fig. 1 Topic-Dependent Recognition based on Topic Detection

computational overhead and also hampers scalability, as the addition of each new topic domain requires an additional recognition process.

One problem in implementing the proposed architecture is that topic detection errors can occur as it is performed on the recognition hypothesis of a single utterance. These errors may propagate through the system, causing an incorrect topic-dependent language model to be selected and thus resulting in highly erroneous recognition hypotheses. Therefore, a mechanism that provides robustness against topic detection errors is required. For this purpose, we introduce a hierarchical back-off mechanism that applies detailed topic models when topic detection is confident and wider models (that cover multiple topics) in cases of uncertainty.

Previous studies have typically investigated topic-dependent recognition on long speech materials such as transcription of news articles [18], [19] and the Switchboard corpus [20]. In these studies, a large number of utterances were used to perform topic detection, thus detection errors were not considered. Also a rescoring framework was typically used, which provided only a limited gain in recognition accuracy while requiring the generation of a large N-best list, which is computationally expensive. In the proposed approach, we re-perform decoding, applying an appropriate topic-dependent language model from the topic detection result in the initial recognition pass. The topic detection result can also be used for other applications, such as improving the back-end translation system performance for speech-to-speech translation [21].

## 2. Language Model Switching Based on Topic Detection

An overview of the proposed system is shown in Figure 1. Speech recognition is performed in two stages. In the first recognition stage, a G-LM (generalized language model) built from the entire training set is applied and topic detection is performed on the recognition result of this pass. Based on the topic detection result and its confidence, an appropriate granularity of TD-LM (topic-dependent language model) is selected. This TD-LM is then used to re-decode the utterance. As a final fallback, the result of the topic dependent recognition and that of the initial topic in-

dependent recognition are compared and the hypothesis with maximum ASR score is selected. (The ASR score is a weighted product of the acoustic and language model probabilities.) This allows the system to back-off completely to the topic-independent G-LM in cases where the TD-LM hypothesis is unlikely. System turn-around time can be reduced by running the current topic-dependent and generalized recognition in parallel and performing re-decoding only when a topic change occurs.

The recognition performance of the proposed architecture is dependent on the TD-LMs applied and the accuracy of topic detection. When TD-LMs cover narrow or individual topics, a large increase in recognition accuracy can be gained, however, topic detection errors will also increase. Training LMs for very narrow topics also generally suffer from data sparseness. On the other hand, LMs that provide coverage over wide topics will reduce the gain in recognition accuracy. In this paper, a multi-layer framework is introduced where a hierarchy of LMs is generated that cover an increasing number of topics. Individual topic LMs are applied when topic detection is confident, and in cases of uncertainty, the system backs-off to wider models that provide coverage over multiple or all topic classes.

### 3. Topic Detection

Topic detection is performed in a manner similar to that used for call-routing [16], [17]. Each sentence in the training set was initially manually labeled with a single topic, from a set of pre-defined topic classes  $T = t_1, \dots, t_M$ . The features used for topic detection consist of word base form tokens (word tokens with no tense information). Appropriate cutoffs are applied to remove those features with low occurrence in the training set. The set of word features for an input sentence is defined by the occurrence counts of each word feature  $w_i$ , that is  $W = (w_1, w_2, \dots, w_V)$  where  $V$  is the vocabulary size.

In this study, we investigate two topic detection methods: unigram likelihood, and SVM (Support Vector Machines). Initially, classification models are trained for each topic class. Topic detection is then performed by applying each classification model to the input utterance and selecting the topic with maximum classification score.

#### 3.1 Unigram Likelihood Based Topic Detection

In this approach, topic-dependent unigram language models are trained for each topic class. The unigram probabilities  $p(w_i|t_j)$  are estimated based on the occurrence counts of each word feature  $w_i$  in the training sentences of that topic  $t_j$ .

The topic classification score ( $score_{UNI}(X, t_j)$ ) is calculated as log-likelihood of topic  $t_j$ 's unigram

model for the input sentence  $X$ , consisting of  $N$  words  $(x_1, \dots, x_N)$ . At recognition,  $X$  is the 1-best hypothesis from the initial recognition pass. The topic detection result is the topic with maximum score.

$$score_{UNI}(X, t_j) = \sum_{i=1}^N \log(p(x_i|t_j)) \quad (1)$$

#### 3.2 SVM Based Topic Detection

SVM (support vector machines) [22] is a popular classification technique based on margin maximization. SVM has been shown to be appropriate for text classification tasks, which typically consist of sparse high-dimensional vector space models. In this approach, a sentence is represented as an individual point within this space, where vector components relate to word token occurrence counts. Based on this vector space model, an SVM hyperplane  $H_j$  is trained for each topic class  $t_j$ . Sentences labeled with that topic ( $t_j$ ) are used as positive examples and the remainder of the training set is used as negative training examples. Due to the high dimensionality of this space, up to 10,000 features, a linear SVM kernel is adequate for classification.

Topic detection is performed by comparing the vector representation of the input sentence ( $X$ ) to each SVM hyperplane. The vector representation  $W = (w_1, \dots, w_V)$  is derived by counting word occurrences in the input sentence  $X = (x_1, \dots, x_N)$ . The topic classification score ( $score_{SVM}(X, t_j)$ ) is calculated as the perpendicular distance between  $W$  and the hyperplane ( $H_j$ ) of topic  $t_j$ . This value should be positive if  $W$  is in-class, and negative otherwise. The detection result is the topic with maximum classification score.

$$score_{SVM}(X, t_j) = dist_{\perp}(W, H_j) \quad (2)$$

### 4. Topic Dependent Language Modeling

The corpus used in this work contains manually assigned topic tags for each sentence. Although these tags can be used directly to train TD-LMs, the resulting models may not be optimal in terms of either perplexity or topic detection accuracy due to subjective definition of labels and inconsistencies between labelers. Thus, each sentence in the training set is re-labeled with the result from topic detection.

In the case of unigram re-labeling, initial unigram models are created based on the original hand-labeled topic tags, and each sentence in the training set is re-labeled as the topic with maximum classification score. This process of model creation and data re-labeling is repeated until convergence. For SVM re-labeling, this process is done only once. Topic detection models are created directly from the hand-labeled tags, and the

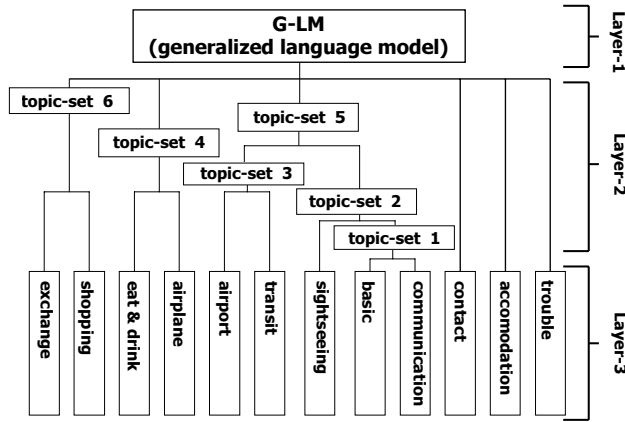


Fig. 2 SVM-based Language Model Hierarchy

training set is then re-labeled using these models. The re-labeling process improves topic detection accuracy and reduces LM perplexity by clustering similar sentences together.

TD-LMs are trained based on the new labels. To reduce the effect of data sparseness, each TD-LM is then linearly interpolated with the domain independent G-LM, which is trained on the entire training set. Interpolation weights are selected to minimize the perplexity of a development set which has been re-labeled in the same manner.

## 5. Language Modeling Based on Hierarchical Topic Classification

To increase the system’s flexibility and robustness, a hierarchical topic back-off mechanism is introduced. In this approach, rather than applying only topic-dependent language models that provide coverage over individual topics, a hierarchy of language models is constructed that provides coverage over an increasing number of topics.

The topic hierarchy is automatically constructed by clustering together those topics likely to be confused during topic detection. The resulting hierarchy for the SVM case is shown in Figure 2. The top node (layer-1) corresponds to a topic-independent G-LM that provides coverage over all topics. The lowest layer (layer-3) corresponds to the most detailed models that provide coverage for individual topics. Intermediate nodes in layer-2 correspond to models that cover multiple topics. Intermediate nodes lower in the hierarchy provide coverage for topics more likely to be confused during topic detection. Ascending the hierarchy, the language models become less topic dependent and typically cover an increasing number of topics.

When the topic detection result is confident, increased recognition accuracy can be gained by applying a language model lower in the hierarchy, which is more topic-dependent. In cases of uncertainty, however, the system should back-off to an intermediate model cov-

Table 1 Topic Hierarchy Construction

$T \leftarrow t_1, t_2, \dots, t_M, k \leftarrow  T $
<b>while</b> $k > 2$ <b>do</b>
determine $t_i, t_j$ such that $dist(t_i, t_j)$ is minimized ( $i \neq j$ )
merge $t_i$ and $t_j$ to create parent node $t_{i,j}$
include $t_{i,j}$ to topic set $T$
remove $t_i$ and $t_j$ from $T$
$k \leftarrow k - 1$
<b>end while</b>

ering multiple plausible topics, or to the topic independent G-LM, rather than selecting a possibly incorrect individual topic. This approach reduces the cases where a topic-dependent language model is selected that does not match the current utterance.

In the following sub-sections, we describe the topic hierarchy clustering algorithm, the inter-topic distance measures used during clustering, and the hierarchy back-off mechanism used to select an appropriate TD-LM at runtime.

### 5.1 Topic Hierarchy Construction

The topic hierarchy is automatically constructed applying agglomerative hierarchical clustering, as described in Table 1. Clustering involves iteratively determining the closest topic pairs and merging them, until only two clusters remain. The resulting hierarchy is then pruned of outlying models. Those models that realize less than a 10% reduction in perplexity compared to the G-LM are removed from the hierarchy. The resulting hierarchy for the SVM case after pruning is shown in Figure 2.

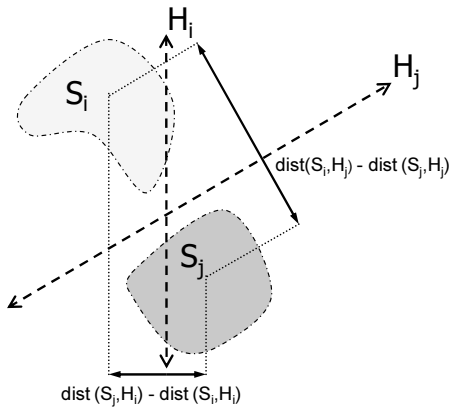
During clustering, an inter-topic distance measure related to topic detection confusability is required. This distance measure is dependent on the topic detection method used.

### 5.2 Unigram-based Inter-topic Distance

For unigram-based topic detection, the distance between two topics  $t_i$  and  $t_j$ ,  $dist_{UNI}(t_i, t_j)$ , is calculated as the normalized log-likelihood based on the training data and unigram models for each topic class (Equation 3.) Normalization is applied to compensate for varied topic class sizes and perplexities.

$$dist_{UNI}(t_i, t_j) = \frac{\sum_{X \in S_i} score_{UNI}(X, t_j)}{\sum_{X \in S_i} score_{UNI}(X, t_i)} + \frac{\sum_{X \in S_j} score_{UNI}(X, t_i)}{\sum_{X \in S_j} score_{UNI}(X, t_j)} \quad (3)$$

$S_i$ : set of training sentences of topic class  $t_i$



**Fig. 3** Inter-topic Distance for SVM-based Topic Detection

### 5.3 SVM-based Inter-topic Distance

For SVM-based topic detection, the inter-topic distance measure is defined as the average distance between topic  $t_i$ 's training set ( $S_i$ ) and topic  $t_j$ 's SVM hyperplane ( $H_j$ ) and vice versa (Equation 4). The distance perpendicular to the SVM hyper plane is used as it directly relates to the topic detection score used.

$$\begin{aligned} dist_{SVM}(t_i, t_j) = & \\ & \left\| \text{average}_{X \in S_i} score_{SVM}(X, t_j) - \text{average}_{X \in S_j} score_{SVM}(X, t_j) \right\| + \\ & \left\| \text{average}_{X \in S_j} score_{SVM}(X, t_i) - \text{average}_{X \in S_i} score_{SVM}(X, t_i) \right\| \end{aligned} \quad (4)$$

$S_i$ : set of training sentences of topic class  $t_i$

A visual representation of this distance measure is given in Figure 3, where

$$dist(S_i, H_j) \equiv \text{average}_{X \in S_i} score_{SVM}(X, t_j)$$

### 5.4 Hierarchical Language Model Back-off

In the proposed recognition architecture, topic detection involves selecting an appropriate LM from the topic hierarchy. This model is then applied during the topic dependent recognition pass. Model selection is based on a hierarchical back-off mechanism which is dependent on the topic detection method used.

For unigram-based topic detection, unigram models are trained for each node in the hierarchy. Topic detection involves selecting the node in all layers that gives the maximum unigram likelihood.

In the SVM case, an individual topic model is used when the SVM classification score for only one topic is positive. Otherwise, we determine the two topics with

**Table 2** Description of Basic Travel Expression Corpus

Training-set: 12 topics, 168,818 sentences
Lexicon size: 18k
Development-set: 10,346 sentences
Test-set: 1,990 utterances (0.67 OOV rate)

the highest topic classification scores, and select their lowest parent node in the hierarchy.

## 6. Experimental Evaluation

### 6.1 Evaluation Corpus

The proposed recognition architecture was evaluated on the ATR Basic Travel Expression Corpus (BTEC) [23]. This corpus consists of Japanese sentences that Japanese travelers are likely to use or encounter while traveling overseas. An overview of the corpus is shown in Table 2. In this study, 12 sub-domain topic classes, as described in Table 3 are used. The original corpus consists of a larger set of 15 topic tags, however, very small topic classes ( $< 2,000$  training sentences) were merged to obtain the above set of topics. For example, the topic classes "restaurant", "snack food", and "drinks" were merged to form the single topic class "eat&drink".

The training set of 168,818 sentences was used for training all language and topic detection models, and for constructing the topic hierarchy. The development set of 10,346 sentences was used to determine the linear interpolation weights applied during TD-LM smoothing. The test set of 1,990 utterances was used for evaluation.

### 6.2 Experimental Setup

Recognition was performed with our Julius recognition engine [24]. For acoustic analysis, 12-dimensional MFCC, energy, and first- and second-order derivatives were computed. The acoustic model applied during recognition was a triphone HMM with 1,841 shared states and 23 Gaussian mixture components set up for 26 phones.

For the baseline ASR system, a topic-independent generalized LM (G-LM) trained on the entire training set was used. On the test set, this baseline model had perplexities of 44.8 (2-gram) and 23.8 (3-gram) and a WER (Word Error Rate) of 8.08%.

### 6.3 Effect of Topic Dependent Language Modeling

First, the effect of topic-dependent language modeling (TD-LM) on test-set perplexity was investigated. We compared three topic labeling schemes: the original hand-labels, re-labeling using unigram topic detection, and SVM-based re-labeling. Based on these labeling schemes TD-LMs were then trained, and the test-set perplexity was calculated (Table 4).

**Table 3** Description of Topic Classes

Topic Class	No. of Training Sentences	Example sentences
accommodation	16473 (10%)	I asked for a room with a shower / What's the price, excluding meals?
airplane	7117 (4%)	Which channel is the film on? / What kind of drinks do you have?
airport	10101 (6%)	Please show me your passport and immigration form / Where can I change travelers cheques?
basic	21880 (13%)	Sorry / Hi / What does that mean? /
communication	11613 (7%)	Are you from New Zealand? / May I use your bathroom?
contact	7514 (4%)	Extension two thirty four please / Call to Japan please / Sorry wrong number
eat&drink	18746 (11%)	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?
shopping	19278 (11%)	How will you pay for this / Can I buy it in Japanese yen?
sightseeing	14581 (9%)	Do you know where the baseball stadium is? / Could you take a photo of us, please?
transit	17027 (10%)	Is this the right platform for the train to Minneapolis?
trouble	22337 (13%)	Can you send a mechanic please? / There's something wrong with the clutch

**Table 4** Perplexities by Topic Dependent Language Modeling

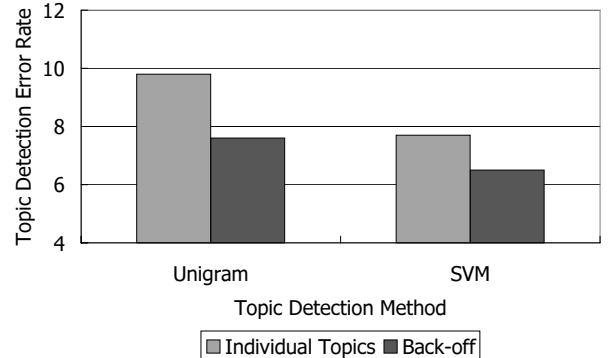
Topic Labeling Method	Perplexity (Reduction compared to G-LM)	
	2-gram	3-gram
G-LM	44.78 (-)	23.77 (-)
Hand	33.51 (25.2%)	18.94 (20.2%)
Unigram	28.00 (37.5%)	16.85 (29.1%)
SVM	29.60 (34.0%)	17.34 (27.1%)

For the baseline G-LM system, perplexities by unigram and tri-gram models were 44.8 and 23.8, respectively. TD-LMs created based on the original hand-labeled topic tags provide a 20.2% reduction in perplexity over the single G-LM. This reduction verifies the effectiveness of topic dependent modeling.

Compared to the hand-labeled case, both unigram and SVM-based re-labeling provide a further reduction in perplexity: 29.1% and 27.1%, respectively. This shows the effectiveness of automatic re-labeling. Unigram re-labeling provides lower perplexity than the SVM case, because there is consistency between the log-likelihood score used for topic detection and the perplexity measure. The unigram method also tends to divide the training set more evenly over the 12 topic classes, while SVM re-labeling results in class sizes similar to the original topics. In the unigram case, the smallest topic class contains 5% of the training data, while in the SVM case this is only 1%.

#### 6.4 Performance of Topic Detection

Next, the performance of the unigram and SVM-based topic detection methods were compared. The re-labeling process was applied to the correct transcriptions of the test set, and the automatically assigned tag is regarded as the "correct" topic of that sentence. Detection accuracy was evaluated by comparing the result when topic detection is applied to the ASR hypothesis to these topic tags. The topic detection error rate for the two methods when only individual topics (layer-1) and when hierarchical back-off (layer-1 or layer-2) was applied are shown in Figure 4. For the hierarchical back-off case, the topic detection result is correct

**Fig. 4** Performance of Topic Detection Methods

if either the correct individual model (layer-3), or its parent model from layer-2 is selected.

Both unigram and SVM-based methods achieve high performance with detection error rates of 9.8% and 7.7%, respectively. SVM significantly outperforms the unigram-based method for both the individual topic case (relative reduction of 21.4%) and for the hierarchical back-off case (relative reduction of 14.6%). This indicates that SVM realizes improved robustness against recognition errors. For both methods, the hierarchical back-off mechanism reduces topic detection errors by around 14%. This shows the effectiveness of the back-off mechanism.

#### 6.5 Performance of Topic Dependent ASR

The proposed recognition framework was implemented by combining topic detection and TD-LMs. The recognition performance (WER) for the unigram and SVM-based systems when various layers of the topic hierarchy are used are shown in Table 5. The baseline system applies only the G-LM (layer-1) during recognition.

For reference, the system performance when "oracle" topic detection is applied is also shown. In this scheme, the correct transcription of the input utterance is used for topic detection instead of the ASR result. Applying this approach when only the layer-3 models are applied, relative reductions in WER of 8.9%, and

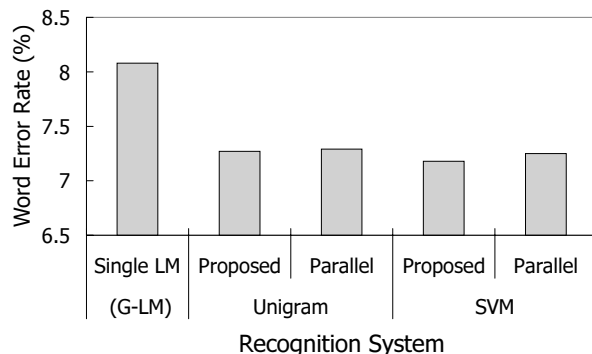
**Table 5** Topic Dependent Recognition Performance

Classification Method	WER % (Relative reduction)			
	Layer 1 (G-LM) baseline	Layer 3	Layers 1,3	All Layers
Topic detection applied to correct transcription (Oracle)				
Unigram (Oracle)	8.08	7.36 (8.9%)	6.93 (21.7%)	6.91 (21.8%)
SVM (Oracle)	8.08	7.64 (5.2%)	7.10 (12.0%)	7.04 (12.0%)
Topic Detection applied to ASR result				
Unigram	8.08	8.12 (-0.5%)	7.36 (8.9%)	7.30 (9.7%)
SVM	8.08	8.24 (-1.2%)	7.42 (8.2%)	7.25 (10.3%)

5.2% are gained over the baseline system for the unigram and SVM systems, respectively. This shows that improved recognition accuracy can be gained by more constrictive, topic-dependent language modeling. Compared to the SVM system, the unigram system had a slightly lower WER. Including the comparison with the layer-1 model (G-LM) further improves recognition accuracy. For around 5% of the utterances, the topic independent G-LM model gave a better recognition hypothesis than the appropriate topic model. As the G-LM is trained over the entire training set, it is less affected by data sparseness than the individual topic models. The inclusion of the layer-2 models using hierarchical back-off provided little improvement in recognition accuracy in the oracle case.

Next, we investigate the system performance when TD-LMs are selected based on the topic detection result from the ASR hypothesis in the initial recognition pass. When only the layer-3 TD-LMs are applied, the recognition performance drops below that of the baseline system. This shows that applying an incorrect TD-LM model significantly degrades recognition performance for that utterance. In the SVM case, even a small number of mis-classified utterances (less than 8%) degrades the overall system performance from the baseline. Introducing the comparison with the layer-1 mitigates the effect of topic detection errors. This comparison is vital for effective performance. The inclusion of the layer-2 models selected by hierarchical back-off further improves recognition accuracy. Compared to the baseline system, a relative improvement of around 10% for both systems is gained. This shows that the proposed hierarchy back-off mechanism realizes robustness against topic detection errors caused in the initial ASR pass.

While the unigram method provided a large reduction in TD-LM perplexity and WER for the oracle case, SVM-based topic detection improved topic detection robustness. Both approaches realize comparable recognition performance when combined with the proposed architecture.

**Fig. 5** Comparison with Parallel Systems

## 6.6 Comparison with Parallel Decoding Scheme

Next, the performance of the proposed framework is compared with a parallel recognition scheme, which performs recognition in parallel with the layer-1 (G-LM) and all layer-3 TD-LMs. The recognition result with maximum ASR score is output. The recognition performance for unigram and SVM-based TD-LMs combined with the proposed framework and parallel decoding is shown in Figure 5.

Both decoding schemes provide a significant reduction in WER compared to the baseline system. Although similar performance is gained with the two approaches, the proposed framework had a much lower computational cost, requiring only 2 recognition processes, compared to 13 for the parallel system.

## 6.7 Extension to Multiple Topics

In the proposed recognition framework, a single TD-LM is selected based on the topic detection result. This framework can be extended to perform recognition with multiple TD-LMs. In this case, the top  $m$  topics with highest topic classification scores are determined by topic detection, and then speech recognition (re-decoding) is performed in parallel with the selected TD-LMs. The recognition results from the G-LM and TD-LMs are finally compared and the hypothesis with maximum ASR score is output. This approach will reduce the number of topic detection errors and improve recognition performance, but incurs increased computational cost compared to the original framework.

The performance of the extended system is shown in Figure 6 for various values of  $m$ . As the number of TD-LMs applied was increased, recognition performance was also improved. At  $m = 3$  the system performance was similar to that using the proposed framework with hierarchical back-off (Proposed). However, at  $m = 6$  the system performance begins to degrade. Since it is not easy to determine the best value of  $m$  a-priori, the proposed framework, applying a single TD-LM, offers a reasonable solution and requires only two recognition passes.

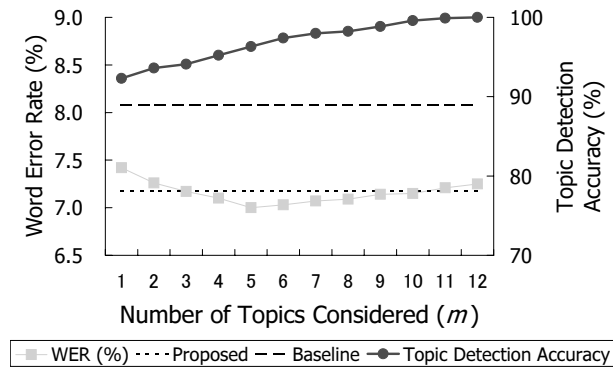


Fig. 6 Extension to Multiple TD-LMs

## 7. Conclusion

We have presented an efficient speech recognition architecture for multi-domain spoken language systems combining topic detection and topic-dependent language modeling. In the proposed approach, the inferred domain of the user's utterance is automatically detected and speech recognition is then performed with an appropriate topic-dependent language model. To improve robustness against topic detection errors, a hierarchical back-off mechanism was introduced that applies detailed topic models when topic detection is confident, and applies wider models that cover multiple topics in cases of uncertainty.

The performance of the proposed architecture was evaluated when combined with unigram and SVM-based topic detection. On the ATR Basic Travel Expression Corpus, both methods provided improved recognition performance compared to a single language model system (relative reductions in WER of 9.8% and 10.3%, respectively). The unigram-based approach provided lower TD-LM perplexity and improved recognition accuracy when the topic was given, however, SVM provided improved topic detection robustness to ASR errors. The overall system performance of both methods was comparable. Finally, in comparison with a parallel approach, the proposed architecture achieves similar recognition accuracy while requiring much less computational cost. Thus, the proposed system realizes more accurate speech recognition in an efficient manner.

## Acknowledgment

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

## 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 Transactions on Speech and Audio Processing*, Vol. 8, No. 1, January 2000.
- [2] S. Seneff and J. Polifroni, "A new restaurant guide conversational system: Issues in rapid prototyping for specialized domains," *Proc. ICSLP*, vol. 2, pp. 665-668, 1996.
- [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 Communication*, vol. 17, pp. 1-18, 1995.
- [4] E. den Os, L. Boves, L. Lamel, and P. Baggia, "Overview of the ARISE project," *Proc. EUROSPEECH*, pp. 1527-1530, 1999.
- [5] R. Carlson, "The dialog component in the WAXHOLM system," *Proc. Eurospeech*, 1996.
- [6] ed. Wolfgang Wahlster, "VERMOBIL: Foundations of speech-to-speech translation," Berlin: Springer, 2000.
- [7] 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. Vaufraydaz, and L. Taddei, "A multi-perspective evaluation of the NESPOLE! speech-to-speech translation system," *Proc. ACL*, pp. -, July, 2002
- [8] 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," *Proc. ICSLP*, pp. 957-960, 1998.
- [9] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue, "Galaxy-II: A reference architecture for conversational system development," *Proc. ICSLP*, pp. 931-934, 1998
- [10] 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, S. Whittaker. "DARPA Communicator dialog travel planning systems: The June 2000 data collection," *Proc. EUROSPEECH*, pp. , 2001.
- [11] M. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, D. Stallard. "DARPA Communicator: Cross-system results for the 2001 evaluation," *Proc. ICSLP*, pp. , 2002.
- [12] D. B. Moran, A. J. Cheyer, L. E. Julia, D. L. Martin, and S. Park, "Multimodal user interfaces in the Open Agent Architecture," *Proceedings of the International Conference on Intelligent User Interfaces* Vol.6-9 pp. 61-70, 1997.
- [13] T. Kawahara, M. Araki, and S. Doshita, "Heuristic search integrating syntactic, semantic and dialog-level constraints," *Proc. IEEE-ICASSP*, Vol.2, pp.25-28, 1994
- [14] F. Wessel and A. Baader, "Robust dialogue-state dependent language modeling using leaving-one-out," *Proc. IEEE-ICASSP*, Vol.2, pp. 741-744, 1999.
- [15] S. Seneff, R. Lau, and J. Polifroni, "Organization, communication, and control in the Galaxy-II conversational system," *Proc. EUROSPEECH*, pp.1271-1274, 1999.
- [16] J. Chu-Carroll and B. Carpenter, "Vector-based natural language call routing," *Computational Linguistics*, vol. 25, no. 3, pp. 361-388, 1999.
- [17] P. Haffner, G. Tur, and J. Wright, "Optimizing SVMs for complex call classification," *Proc. ICASSP-2003*, vol. 1, pp. 632-635, 2003.

- [18] S. C. Martin, J. Liermann, and H. Ney, "Adaptive topic-dependent language modelling using word-based variograms," *Proc. EUROSPEECH*, pp. 1447–1450, 1997.
- [19] K. Seymore and R. Rosenfeld, "Using story topics for language model adaptation," *Proc. EUROSPEECH*, pp. 1987–1990, 1997.
- [20] S. Khudanpur and J. Wu, "A maximum entropy language model integrating n-grams and topic dependencies for conversational speech recognition," *Proc. IEEE-ICASSP*, Vol.1, pp. 553–556, 1999.
- [21] M. Paul, E. Sumita, and S. Yamamoto, "Topic-dependent word selection," *Proc. FIT*, Japan, pp. 77–78, 2002.
- [22] T. Joachims, "Text categorization with Support Vector Machines: learning with many relevant features," *Proc. ECML*, pp. 137–142, 1998.
- [23] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Towards a broad-coverage bilingual corpus for speech translation on conversations in the real world," *Proc. LREC*, pp. 147–152, 2002.
- [24] A. Lee, T. Kawahara, and K. Shikano, "Julius— an open source real-time large vocabulary recognition engine." *Proc. EUROSPEECH*, pp. 1691–1694, 2001.