LANGUAGE MODEL SWITCHING BASED ON TOPIC DETECTION FOR DIALOG SPEECH RECOGNITION

Ian R. Lane[†‡], Tatsuya Kawahara[†‡], and Tomoko Matsui[‡]

[†]School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
[‡]ATR Spoken Language Translation Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

ABSTRACT

An efficient, scalable speech recognition architecture is proposed for multidomain dialog systems by 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 domain-dependent language model. The architecture improves accuracy and efficiency over current approaches and is scaleable to a large number of domains. In this paper, a novel framework using a multi-layer hierarchy of language models is introduced in order to improve robustness against topic detection errors. The proposed system provides a relative reduction in WER of 10.5% over a single language model system. Furthermore it achieves an accuracy that is comparable to using multiple language models in parallel while using only a fraction of the computational cost.

1. INTRODUCTION

In recent years, there has been a large growth in the development and public use of telephone-based spoken dialog systems. One area that is now of interest is providing increased usability by allowing users to access information from multiple domains [1]. When performing speech recognition over multiple domains, topic- or sub-task-dependent language modeling increases both the accuracy and efficiency of the system. This approach is also convenient for development modularity, as new domains can be added to the system without affecting the accuracy of the existing domains.

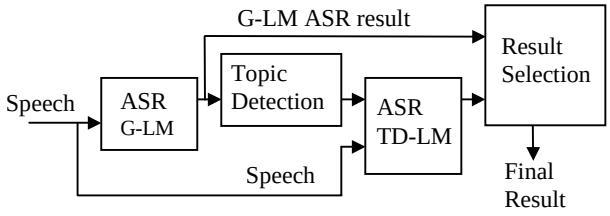
Current dialog systems that use multiple TD-LMs (topic-dependent language models) for recognition mainly adopt a system initiative approach [2]. These systems prompt the user and apply an appropriate LM based on the internal state of the system. Such systems do not allow any user initiative and thus have low usability. Increased usability can be achieved by allowing users to switch between domains, but in most cases, users still must explicitly state the domain they require before they can query that domain [1].

In call routing systems [3], the topic of the user's initial utterance is implicitly detected by performing topic detection on the recognition result. A similar technique can be used for dialog systems to automatically determine the domain required, and as utterances in the same topic are likely to follow, applying a topic-dependent LM is advantageous.

In the proposed system, a combination of topic detection and topic-dependent LMs are used to allow the user to seamlessly switch between domains while maintaining high recognition accuracy. One problem in implementing this architecture is that errors can occur as topic detection is performed based on a single utterance. A mechanism that provides robustness against topic detection errors is required.

Previous studies have typically investigated topic-based recognition on long speech materials such as the transcription of news articles and the Switchboard corpus [4]. In these studies, a large number of utterances were used to perform topic detection, and thus topic detection errors were not considered. A rescoring framework was also used that provided only a limited gain in recognition accuracy while requiring the generation of a large N-best list, which is computationally expensive. Decoding with multiple TD-LMs in parallel is another possible solution, but requires large computational overhead. The parallel approach also offers little scalability, as the addition of each new topic domain requires an extra recognition process.

This paper proposes a method that re-performs decoding based on the topic detected in the initial recognition pass. This approach uses an appropriate TD-LM for recognition and thus provides recognition gain with moderate computational overhead. A novel framework using multi-layer hierarchical LMs is introduced in order to provide increased robustness in cases where topic detection is difficult or erroneous.



G-LM: Generalized Language Model
TD-LM: Topic-dependent Language Model

Figure 1: System Architecture

2. SYSTEM OVERVIEW

An overview of the proposed system architecture is shown in Figure 1. The initial recognition is performed with a G-LM (generalized language model) built from the entire training set. This model covers all topics and can thus be used to perform topic detection.

Topic detection based on the result of the initial recognition pass is performed using unigram models for each topic. The LM of the topic with the maximum log-likelihood is then used to re-decode the utterance. The system turn-around time can be minimized by running the current topic-dependent and generalized recognition in parallel and performing re-decoding only when a topic change occurs. Since topic detection is performed based on a recognition hypothesis, topic detection errors may occur and propagate through the system. This would cause an incorrect topic LM to be selected for decoding, and thus the result would likely contain many recognition errors.

In the final stage, the initial hypothesis using the G-LM is compared with the result of topic-dependent decoding. ASR-score is used to select the best hypothesis. This fallback mechanism is used to correct errors incurred in topic detection. In this process, the system reverts to the original G-LM result if the topic-dependent result seems unlikely.

The interaction between the TD-LMs used and the topic detection accuracy is important for the performance of this architecture. When TD-LMs cover narrow topics, a large increase in recognition accuracy can be gained, but the topic detection accuracy declines. Training LMs for very narrow topics also generally suffer from data sparseness. Topic detection of wider topics is more accurate, but the gain in recognition accuracy is reduced.

In the approach described in this paper, a multi-layer framework is proposed where a hierarchy of LMs is generated that cover an increasing number of topics. This allows the use of narrow topic LMs when topic detection is confident and wider topics in cases of uncertainty. The top node corresponds to the G-LM and is used as the last fallback.

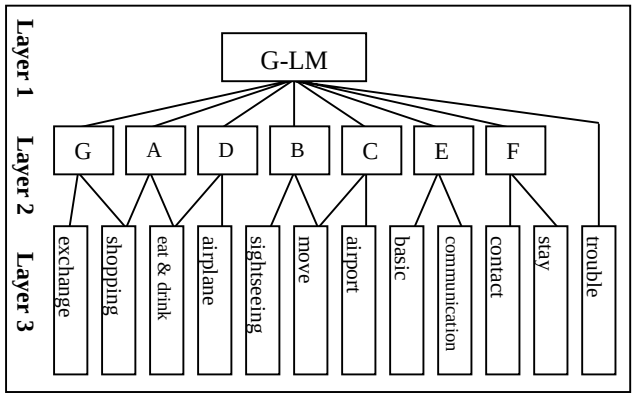


Figure 2: Multi-layer Language Modeling

3. TOPIC DEPENDENT LANGUAGE MODELING

3.1. Training Set Clustering

The corpus used for evaluation had topic labels that were manually assigned. Using these to produce LMs would not necessarily be optimal in terms of perplexity. Thus an iterative re-labeling method is applied where each sentence in the training set is re-labeled with the topic that gives minimum perplexity. Initial unigram models are created for each topic based on the original hand-labeled topic tags. This process of unigram creation and re-labeling is repeated until convergence.

The re-labeling process reduces the LM perplexity by clustering similar sentences together. This in effect narrows the topic of each of the clusters and better models utterances within that topic. Each TD-LM is then linearly interpolated with the G-LM to reduce the effect of data sparseness. Interpolation weights are selected to minimize the perplexity of the development set.

3.2. Multi-layer Language Modeling

To increase the system's flexibility and robustness, a hierarchical LM framework is introduced. An intermediate layer of language models is created to cover multiple topics. These topics can be detected more reliably than in the individual topic case. An example multi-layer hierarchy constructed with the experiment corpus is shown in Figure 2. The top node corresponds to a topic-independent G-LM that gives complete coverage of all topics, and the bottom layer corresponds to the most detailed, individual topic models.

The construction of the multi-layer hierarchy involves creating a set of intermediate nodes that cover multiple topics. These nodes are created by merging the topic pairs with minimum normalized cross-perplexity (Eq. (1)). This procedure is described in detail in Figure 3. In this set of experiments, a single layer of intermediate nodes was created. Here, a CUTOFF of 9 was used. The

for each topic C_i
  determine C_j where DIST(C_i,C_j) is minimal (i≠j)
  if DIST(C_i,C_j) < CUTOFF **then**
    create parent node by merging topics C_i and C_j
  else
    no parent node is created (*topic C_i is unique*)

Figure 3: Multi-layer hierarchy construction algorithm

intermediate nodes are created from a combination of the most complementary pairs, and the topics below them are those most likely to be confused during topic detection. Traversing down the hierarchy both the topic dependency of the LMs and the likelihood of topic detection errors increase.

$$DIST(C_i, C_j) = \frac{PP(T_{Ci}, M_{Cj})}{PP(T_{Ci}, M_{Ci})} + \frac{PP(T_{Cj}, M_{Ci})}{PP(T_{Cj}, M_{Cj})} \quad (1)$$

T_{Ci} : Training set of topic class C_i
M_{Cj}: Unigram model of topic class C_j
$PP(T_{Ci},M_{Cj})$: Perplexity of model M_{Cj} given training set T_{Ci}
$DIST(C_i,C_j)$: Normalized cross perplexity of topics C_i and C_j

In a preprocessing step, topic domains with insufficient data (< 3000 sentences) are merged with the closest larger topic before clustering. Here, normalized cross-perplexity (Eq. (1)) is again used as the distance measure.

4. TOPIC DETECTION

Unigram topic models are created for each node in layers 2 and 3 of the topic hierarchy. Topic detection is performed by calculating the log-likelihood of each of the topic models against the 1-best hypothesis from the baseline recognition pass. The detection result is the topic with the maximum log-likelihood value. In this set of experiments, using the N-best hypotheses to perform topic detection did not improve detection performance.

5. EXPERIMENTAL EVALUATION

The ATR phrasebook corpus [5] was used to investigate the performance of the proposed system. Details of the corpus are given in Table 1.

Recognition was performed with the Julius recognition engine. For acoustic analysis, 12-dimensional MFCC with first- and second-order derivatives are computed. The acoustic model is a triphone HMM with 1841 shared states and 23 Gaussian mixture components set up for 26 phones.

Language: Japanese
Domain: Overseas Travel
Training-set: 12 topics,168818 sentences
Lexicon size: 18k
Development-set: 10346 sentences
Test-set: 1990 utterances (0.67 OOV)

Table 1: Corpus Description

Method	Perplexity (Reduction over G-LM %)	
	2-gram	3-gram
Single G-LM	44.78	23.77
12 topics (hand-labeled)	33.51 (25.2%)	18.94 (20.2%)
12 topics (clustered)	28.00 (37.5%)	16.85 (29.1%)

Table 2: TD-LM perplexities

For the baseline ASR system, a generalized LM (G-LM) trained on the entire training set is used for recognition. On the testset, this baseline LM has perplexities of 44.78 (2-gram) and 23.77 (3-gram). The WER is 8.54%.

5.1. Effect of Topic Dependent Language Model

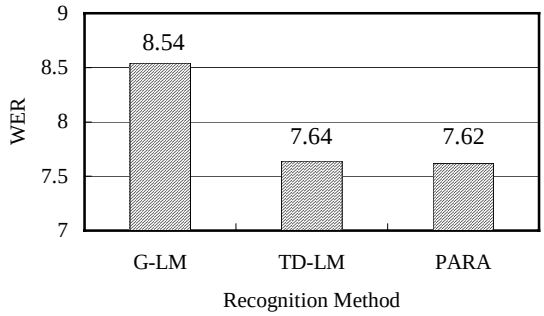
First, topic dependent language models (TD-LM) are created based on the original hand-labeled topic tags. TD-LMs provide a 20.2% reduction in perplexity over a single G-LM (Table 2). This reduction verifies the effectiveness of topic dependent modeling. Next, iterative automated clustering is applied. This further reduces the perplexity (29.1%). This demonstrates that automatic labeling is more effective than the hand-labels.

5.2. Combining Topic Detection with Multi-layer LMs

Recognition is performed in two stages as described in section 2. In the initial recognition pass, recognition is performed with the layer 1 G-LM. Topic detection is performed on the initial ASR hypothesis, and its result is then used to select the appropriate layer 3 (most detailed topic dependent) model or layer 2 (intermediate) model to re-perform decoding. The recognition performance for different sets of topic layers is given in Table 3. As the layer 1 model (G-LM) is always applied in the first recognition pass for topic detection, it is also compared with other models based on the ASR-score. For reference, an oracle method that uses the correct transcriptions for topic detection is also presented.

Classification Method	WER % (relative reduction) (G-LM) Layer 1	Layer 3	Layer 1,3	Layer 1,2,3
Topic Known (Oracle)	8.54	7.68 (10.0%)	6.96 (18.5%)	6.9% (19.0%)
ASR Based Topic Detection	8.54	8.43 (1.3%)	7.81 (8.5%)	7.64 (10.5%)

Table 3: Multi-layer language model performance



G-LM: Single Generalized LM (baseline)
TD-LM: Topic dependent model selected using ASR based topic detection
PARA: All 20 LM in parallel

Figure 5: Comparison of ASR performance

In case of oracle topic detection (Table 3, first row), the most detailed layer 3 models provide a gain of 10% over the baseline system. By including the comparison with the layer 1 model, this gain is increased to 18.5%. For around 5% of the utterances, the layer 1 model gave a better recognition hypothesis than the appropriate topic model. This is because the topic-independent layer 1 model (G-LM) is trained over the entire training set, and is thus less affected by data sparseness than the detailed topic models. Inclusion of the intermediate layer 2 models provides little improvement in this case.

When ASR-based topic detection is performed using only layer 3, the topic detection accuracy is 90.2%. In this case, there is little improvement in WER over the baseline. Introducing the comparison with layer-1 (G-LM) mitigates the effect of topic detection errors, and WER is reduced by 8.5% relatively. By adding the layer 2 models, the accuracy of selecting a topic or a parent node is 92.4%. This method provides a 10.5% relative reduction in WER over the baseline system. The reduction in WER is directly attributed to the improved robustness offered by the multi-layer approach. Utterances that may otherwise be incorrectly detected are referred to a more generalized LM that covers multiple topics, and thus the likelihood of using an incorrect TD-LM for recognition is reduced.

5.3. Comparison with Multiple LMs in Parallel

Finally, the proposed system is compared with a parallel system where recognition is performed in parallel using all 20 LMs and the result with the maximum ASR-score is output. A comparison of the baseline system, the proposed method and the parallel system is shown in Figure 5.

Both the parallel system and the proposed method realize a WER reduction of around 10.5% over the baseline system. While the two methods achieve comparable recognition performance, the computational cost of the proposed method is only about 10% of that of the parallel system.

6. CONCLUSIONS

We have presented an efficient speech recognition architecture based on topic detection and topic-dependent language modeling. The proposed system provides a 10.5% relative reduction in WER over a single LM and achieves recognition performance that is comparable to running a large number of LMs in parallel while requiring a much smaller computational overhead. A novel framework of multi-layer LMs, which is automatically derived without a priori knowledge, is also proposed. This framework provides robustness against topic detection errors, and is critical in achieving improved accuracy while maintaining efficiency.

ACKNOWLEDGEMENTS: 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.” The authors are grateful to Dr. S. Nakamura and Prof. H. G. Okuno of Kyoto Univ. for their comments.

7. REFERENCES

[1] S. Seneff, R. Lau, and J. Polifroni “Organization, Communication, and Control in the Galaxy-II Conversational System”, Proc. Eurospeech, 1999

[2] F. Wessel, and A. Baader, “Robust Dialogue-State Dependent Language Modeling using Leaving-One-Out”, Proc. ICASSP Vol. 2, pp. 741-744, 1999

[3] G. Riccardi, A. Gorin, A. Ljolje, M. Riley, “A spoken Language System for Automated Call Routing”, Proc. ICASSP, Vol. 2, pp. 1143-1146, 1997

[4] S. Khudanpur and J. Wu, “A Maximum Entropy Language Model Integrating N-grams and Topic Dependencies for Conversational Speech Recognition,” Proc. ICASSP ‘99, pp. 553-556, 1999.

[5] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, “Towards a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World”, LREC 2002, pp. 147-152, 2002