Hierarchical Topic Classification for Dialog Speech Recognition based on Language Model Switching

Ian R. Lane[†‡], *Tatsuya Kawahara*[†‡], *Tomoko Matsui*[‡], *Satoshi Nakamura*[‡]

†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

A speech recognition architecture combining topic detection and topic-dependent language modeling is proposed. In this architecture, a hierarchical back-off mechanism is introduced to improve system robustness. Detailed topic models are applied when topic detection is confident, and wider models that cover multiple topics are applied in cases of uncertainty. In this paper, two topic detection methods are evaluated for the architecture: unigram likelihood and SVM (Support Vector Machine). On the ATR Basic Travel Expression corpus, both topic detection methods provide a comparable reduction in WER of 10.0% and 11.1% respectively over a single language model system. Finally the proposed re-decoding approach is compared with an equivalent system based on re-scoring. It is shown that re-decoding is vital to provide optimal recognition performance.

1. Introduction

For improved usability, spoken dialog systems should allow users to retrieve information from multiple domains naturally and efficiently. 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 dialog systems that use multiple TD-LMs (topic-dependent language models) mainly adopt a system initiative approach [1]. These systems prompt the user and apply an appropriate LM based on the internal state of the system. Increased usability can be achieved by allowing users to switch between domains, but in most cases, users still must explicitly state the required domain before they can make a domain dependent query [2]. Decoding with multiple TD-LMs in parallel is one possible solution, but requires large computational overhead. This approach also hampers scalability, as the addition of a new topic domain requires an extra recognition process.

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 TD-LM. This allows the user to seamlessly switch between domains while maintaining high recognition accuracy. As utterances in the same topic are likely to follow, switching between topic-dependent LMs is advantageous.

One problem in implementing this architecture is that topic detection errors may occur as topic detection is performed on the recognition hypothesis of a single utterance. Previous studies typically investigated topic-based recognition on long speech materials such as transcription of news articles and the Switchboard corpus [3]. In these studies, a large number of utterances were used to perform topic detection, thus topic detection errors were not considered. Typically a rescoring framework was used, which was found to provide only a limited gain in recognition accuracy while requiring the generation of a large N-best list, which is computationally expensive. The proposed method re-performs decoding applying an appropriate TD-LM detected in the initial recognition pass. To provide robustness against topic detection errors, 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.

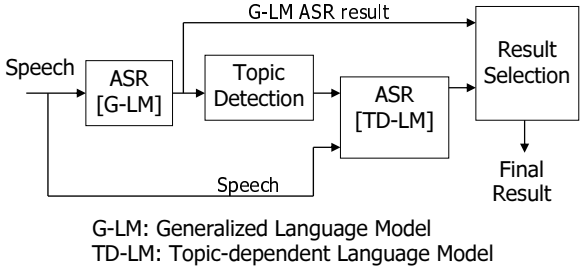

Figure 1: Proposed multi-domain dialog system

An overview of the proposed system is shown in Figure 1. 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. Based on the topic detection result and its confidence, an appropriate granularity of topic model is selected. The selected model is then used to re-decode the utterance. As a final fall-back, the result of the topic dependent pass and that from the initial topic independent pass are compared and the hypothesis with maximum ASR confidence is selected. This allows the system to back-off completely to the topic independent G-LM. 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.

In this paper, we investigate two methods of topic detection: unigram likelihood, and SVM (Support Vector Machines) [4]. These methods are evaluated for the proposed recognition architecture with respect to topic clustering effectiveness, topic detection accuracy and system recognition performance. Finally the proposed re-decoding based approach is compared with an equivalent system based on re-scoring.

2. Topic Detection

In this paper, two topic detection methods are investigated. The first based on unigram likelihood and the second based on SVM. In both methods, topic detection models are trained for each topic using the same data used to create the TD-LMs.

2.1. Unigram Likelihood based Topic Detection

In this method, unigram topic models are created for each topic. Topic detection is performed by calculating the log-likelihood of each topic model against the 1-best hypothesis from the initial recognition pass. The detection result is the topic with maximum log-likelihood value.

2.2. SVM based Topic Detection

Based on a vector space model, each sentence S_i is represented as a point in an n-dimensional vector space $(O(w_1), O(w_2), \cdots, O(w_n))$, where $O(w_k)$ is the number of occurrences of word w_k in S_i. The vector components relate to all words that occur more than once in the training set. The use of a stop-list was not effective in improving the system performance, and is not used here. SVM models are trained for each topic. 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 training examples.

Topic detection is performed by feeding the vector representation of the input sentence to each SVM classifier. The perpendicular distance between the sentence S_i and each SVM hyper-plane is used as a confidence measure for detection. This value is positive if S_i is in-class, and negative otherwise. The detection result is that topic with the maximum confidence score.

3. Topic Dependent Language Modeling

3.1. Topic Re-labeling

The corpus used in this work contains manually assigned topic tags for each sentence. Using these tags to produce TD-LMs is not optimal in terms of either perplexity or topic detection accuracy. Thus re-labeling is applied where each sentence in the training set is re-labeled with the topic given by the topic detection model.

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 minimum perplexity. This process of topic model creation and data re-labeling is repeated until convergence. For SVM re-labeling, this process is only done once. The topic models are created from the hand-labeled data.

This re-labeling process maximizes the topic detection accuracy and reduces the LM perplexity by clustering similar sentences together. This in effect narrows the topic of each of the clusters and thus 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. Language Model Hierarchy

To increase the system's flexibility and robustness, a topic back-off mechanism is introduced. A topic hierarchy is automatically constructed by clustering together those topics likely to be confused during topic detection. An example topic hierarchy based on SVM topic detection is shown in Figure 2. The top node corresponds to a topic-independent G-LM that gives complete coverage of all topics, the bottom layer corresponds to the most detailed, individual topic models, and the intermediate nodes corresponds to models that cover multiple topics.

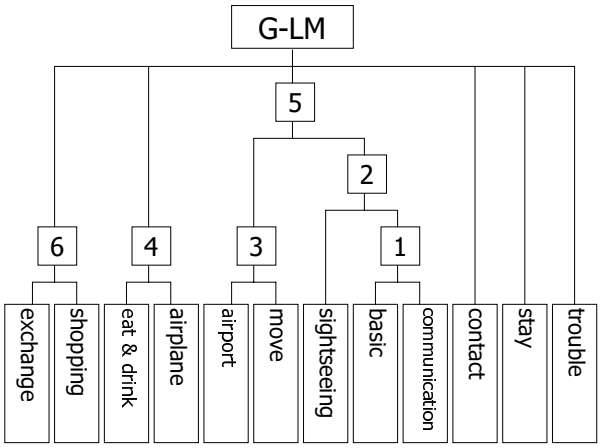

Figure 2: SVM Based Language Model Hierarchy

Models in higher layers cover increasing number of topics, and thus become less topic dependent. When topic detection is confident, individual topic models should be used as they offer the greatest recognition accuracy. In cases of uncertainty, however, the system should back-off to an intermediate model covering multiple plausible topics, or to the topic independent G-LM, rather than selecting a possibly incorrect individual topic. To construct the topic hierarchy, closely related topics are clustered together. A distance measure between topics corresponding to the topic detection method is used. It is defined in a different way for unigram and SVM based topic detection.

3.3. Unigram based Inter-Topic Distance Measure

For unigram based topic detection, the distance between two topics is calculated as the normalized cross-perplexity between the topics (Equation 1). Normalization is required as there are large differences in perplexity between the topic sets.

$$DIST_{UNI}(C_i, C_j) = \frac{PP(T_{C_i}, M_{C_j})}{PP(T_{C_i}, M_{C_i})} + \frac{PP(T_{C_j}, M_{C_i})}{PP(T_{C_j}, M_{C_j})} \quad (1)$$

T_{C_i}: Training set of topic class C_i
M_{C_j}: Unigram topic model for topic class C_j
$PP(T_{C_i}, M_{C_j})$: Perplexity of T_{C_i} by M_{C_j}

3.4. SVM based Inter-Topic Distance Measure

For SVM based topic detection, the distance between two topics C_i and C_j (Equation 3) is calculated as the average distance between topic C_i's training set (T_{C_i}) and C_j's SVM separating hyperplane (X_{C_j}) (Equation 2) and vice versa. The distance perpendicular to the SVM hyperplane is used as it relates to the probability of occurrence of topic detection errors.

Table 1: Basic Travel Expression Corpus Description

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)

$$dist_{avg}(T_{C_i}, X_{C_j}) = \frac{1}{n_i} \sum_{k=0}^{n_i} dist(x_k, X_{C_j}) \quad (2)$$

$$DIST_{SVM}(C_i, C_j) = \| dist_{avg}(T_{C_i}, X_{C_j}) - dist_{avg}(T_{C_j}, X_{C_j}) \| + \| dist_{avg}(T_{C_j}, X_{C_i}) - dist_{avg}(T_{C_i}, X_{C_i}) \| \quad (3)$$

T_{C_i} : Training set of topic class C_i
X_{C_j} : SVM hyperplane for topic class C_j
$dist(x_i, X_{C_j})$: perpendicular distance from SVM hyperplane X_{C_j} to a sample x_i: positive when in-class, negative otherwise
n_i : training set size of topic class C_i

3.5. Topic Hierarchy Construction

Based on the above inter-topic measures, a topic hierarchy is automatically constructed using agglomerative hierarchical clustering. Clustering involves iteratively determining the closest topic pairs and merging them, until only two clusters remain. These two clusters become the direct children of the topic independent G-LM. Finally the resulting hierarchy is pruned of outlying models. Here models that provide less than a 10% reduction in perplexity compared to the G-LM are removed from the hierarchy. The resulting hierarchy for the SVM case is shown in Figure 2.

3.6. Hierarchical Back-Off Mechanism

When using the hierarchical back-off mechanism, topic detection involves selecting an appropriate LM from the hierarchy to be applied in the topic dependent recognition pass. For unigram based topic detection, we create unigram models for each node in the hierarchy. Topic detection simply involves selecting the node with the maximum unigram likelihood. In the SVM case, we select an individual topic model when the SVM score for only one topic is positive. Otherwise we select the parent node of the best two topics.

4. Experimental Evaluation

The ATR Basic Travel Expression corpus [5] was used to evaluate the proposed system. Details of the corpus are given in Table 1. Recognition was performed with our 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.

For the baseline ASR system, a generalized LM (G-LM) trained on the entire training set is used. On the test-set, this baseline LM has perplexities of 44.78 (2-gram) and 23.77 (3-gram). The WER is 8.08%.

Table 2: Perplexities by Topic Dependent Language Models

Method	Perplexity (Reduction over 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%)

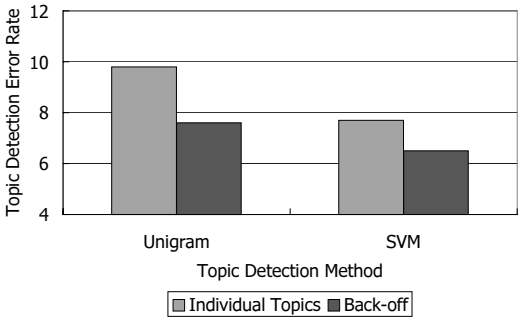

Figure 3: Topic Detection Performance

4.1. Topic Dependent Language Modeling

First the perplexity by topic dependent language modeling is investigated. TD-LMs created based on the original hand-labeled topic tags provide a 20.2% reduction in perplexity over a single G-LM (Table 2). This reduction verifies the effectiveness of topic dependent modeling. Next, re-labeling using unigram and SVM is applied. Both these methods provide a significant reduction in perplexity, 29.1% and 27.1% respectively. This shows the effectiveness of automatic re-labeling. The unigram method is based on term frequency and tends to divide the training set evenly over the 12 topics. For the SVM based method, the resulting topics relate better to the original topic concepts, but cluster sizes are not balanced.

4.2. Topic Detection

Next, we investigate the performance of the two topic detection methods. The topic detection accuracy is evaluated by comparing the ASR based topic detection result with that based on the original transcription. The topic detection error rate with unigram and SVM methods when using only individual topics and the proposed hierarchical back-off case is shown in Figure 3.

SVM based topic detection offers a significant reduction in topic detection errors when compared to the unigram approach: 21.4% for the individual topic case and 14.5% for the topic back-off case. SVM offers improved topic detection robustness against recognition errors. When the topic back-off mechanism is applied, topic detection errors are reduced by 14.5% and 14.3% respectively for the unigram and SVM cases. This shows the effectiveness of the proposed mechanism to reduce topic detection errors.

4.3. Topic Dependent Recognition

Next, the speech recognition performance of the proposed system is investigated. Recognition is performed in two stages as described in section 2. The word error rate of the baseline system using a single G-LM, and the proposed architecture based on unigram and SVM topic detection is shown in Figure 4. The

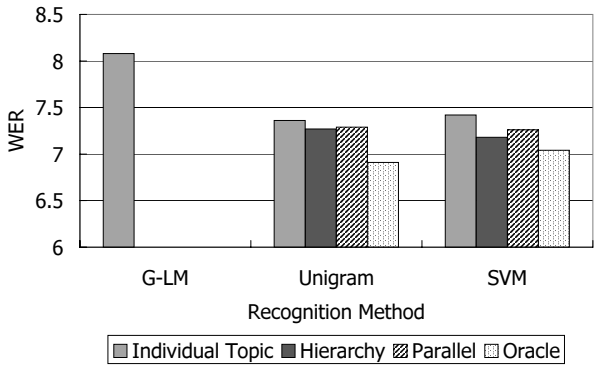

Figure 4: Speech Recognition Performance

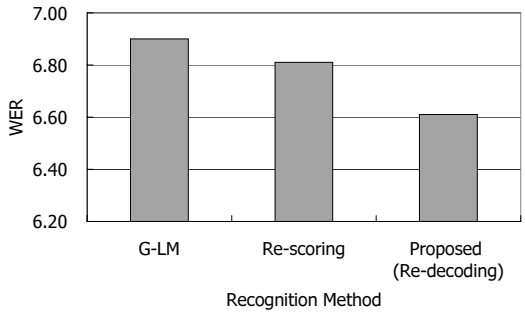

Figure 5: Comparison with Re-scoring and Parallel Systems

performance when using only individual topic models and when using the proposed back-off mechanism is shown. For comparison, the performance of an oracle system that selects the correct topic model and a parallel system is also shown. In the parallel system, recognition is performed with the G-LM and all TD-LMs in parallel and the recognition result with the maximum ASR score is output.

In the case of the oracle system, TD-LMs created with both unigram and SVM re-labeling offer a significant reduction in WER over the baseline system: 14.5% and 12.9% respectively. When the correct topic is known, the unigram approach provides a slight improvement in recognition accuracy over the SVM based system. In the unigram approach the classes are more evenly balanced than in the SVM case and thus language models may be trained more reliably.

Next, the recognition performance of the proposed system is investigated. When only individual topics are used, the WER for the unigram and SVM based systems are 7.36% and 7.42% respectively. Applying the proposed back-off mechanism, the WER is reduced to 7.27% and 7.18%, a relative reduction of 10.0% and 11.1% over the baseline system. This reduction shows the effectiveness of the proposed back-off mechanism. Compared with the parallel system, the proposed architecture offers comparable recognition performance in both the unigram and SVM cases. However, the proposed system applies just two recognition processes while the parallel system applies $n + 1$, where n is the number of topics.

4.4. Comparison with Re-scoring and Parallel Systems

Finally the proposed system, which is based on a re-decoding approach, is compared to a re-scoring based system. Here SVM based topic detection is used and no back-off mechanism is applied. The re-scoring approach involves generating a 1000-best list in the initial recognition pass, and re-scoring this list using the appropriate TD-LM selected by topic detection. In this set of experiments, much wider search parameters are required to generate a large N-best list. This improves system accuracy over the previous experiments, however the recognition time is also much increased.

The performance of the baseline system, and the proposed system applying re-scoring and re-decoding is shown in Figure 5. The re-scoring based approach offers only a slight reduction in WER (1.8%) over the baseline system. Applying the proposed re-decoding based approach a relative reduction in WER of 4.2% over the baseline system, and 2.9% over the re-scoring based method is gained. Applying the proposed back-off mechanism further reduces the WER to 6.54%, a reduction in WER of 5.2% over the baseline system. From these results, it is shown that to achieve significant improvement in performance re-decoding is vital.

5. Conclusion

We have presented an efficient speech recognition architecture combining topic detection and topic-dependent language modeling. In this paper we evaluated two topic detection methods for this architecture: unigram likelihood and SVM. The unigram method was found to offer improved clustering effectiveness. It gave a reduction in TD-LM perplexity and WER for the oracle case over the SVM approach. However, SVM based topic detection provides improved topic detection robustness. Both approaches offer comparable speech recognition performance when used with the proposed architecture. Relative reductions in WER of 10% or more over a single model system were achieved. Finally the proposed system is compared to an equivalent system based on re-scoring, and it is shown that re-decoding is vital for optimal system performance.

Acknowledgments: 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".

6. References

[1] Wessel, F., and Baader, A. "Robust Dialogue-State Dependent Language Modeling using Leaving-One-Out", Proc. ICASSP'99 Vol. 2, pp. 741-744, 1999.

[2] Rudnicky, A. I., Polifroni, Thayer, E. H., and Brenan, R. A. "Organization, Communication, and Control in the Galaxy-II Conversational System", Proc. EUROSPEECH'99, 1999.

[3] Khudanpur, S. and Wu, J. "A Maximum Entropy Language Model Integrating N-Grams and Topic Dependencies for Conversational Speech Recognition", Proc. ICASSP'99, pp. 553-556,1999.

[4] Joachims, T. "Text Categorization with Support Vector Machines", Proc. European Conference on Machine Learning, 1998.

[5] Takezawa, T., Sumita, M., Sugaya, F., Yamamoto, H., Yamamoto S. "Towards a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World", Proc. LREC'02, pp. 147-152,2002