音声言語情報処理　44－25
(2002. 12. 19)

話題同定に基づく言語モデル切替えによる対話音声認識

Lane Ian R. [†‡]　河原 達也 [†‡]　松井 知子 [‡]　中村 哲 [‡]

† 京都大学情報学研究科
〒606-8501 京都市左京区吉田本町

‡ ATR 音声言語コミュニケーション研究所
〒619-0288 京都府「けいはんな学研都市」光台二丁目 2-2

E-mail: † {ian,kawahara}@kuis.kyoto-u.ac.jp, ‡ {tomoko.matsui, satoshi.nakamura}@atr.co.jp

あらまし　複数ドメインの対話システムを構築するために話題同定と話題依存の言語モデルを用いた音声認識手法を提案する。本手法では、ユーザの発話（初期認識結果）から話題を自動的に検出し、その話題に依存した言語モデルを用いて（再度）音声認識を行う。これにより、ドメイン数に関係なく効率性を維持しながら、認識精度の改善を実現する。本稿では、unigram の尤度と SVM に基づく話題同定法を実装・比較する。さらに、話題同定誤りに頑健に対処するため、階層的な言語モデルの枠組みを提案する。本手法により、単一の言語モデルに比べて、単語誤り率が 10.3%改善され、複数の言語モデルを並列に用いた場合と比べて、はるかに少ない計算量で同等の認識精度を得ることができた。

キーワード　音声認識, 対話音声, 話題同定, サポートベクトルマシン, 複数ドメインの対話システム

Language Model Switching Based on Topic Detection for Dialog Speech Recognition

Ian R. LANE [†‡]　Tatsuya KAWAHARA [†‡]　Tomoko MATSUI [‡]　and　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

E-mail: † {ian,kawahara}@kuis.kyoto-u.ac.jp, ‡ {tomoko.matsui, satoshi.nakamura}@atr.co.jp

Abstract　An efficient, scalable speech recognition architecture is proposed for multi-domain 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, unigram likelihood and SVM based topic detection methods are compared. A novel framework using a multi-layer hierarchy of language models is also introduced in order to improve robustness against topic detection errors. The proposed system provides a relative reduction in WER of 10.3% over a single language model system. Furthermore, it achieves an accuracy that is comparable to using multiple language models in parallel while requiring only a fraction of the computational cost.

Keyword　Speech Recognition, Dialog Speech, Topic Detection, Support Vector Machines, Multi-domain Dialog Systems

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.

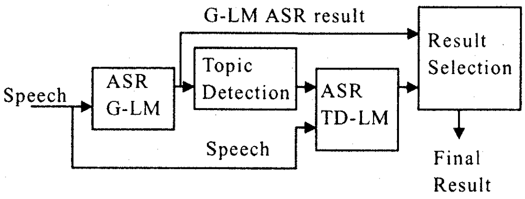
Current dialog systems that use multiple TD-LMs (topic-dependent language models) for recognition mainly adopt a system initiative approach [2]. These systems prompt the user and apply an appropriate LM based on the internal state of the system. Such systems do not allow any user initiative and thus 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 [4,5]. 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 a multi-layer hierarchy of 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 is performed based on the result of the initial recognition pass. The LM of the topic selected 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 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. Such errors would cause an incorrect topic LM to be selected for decoding, and thus the ASR result would likely contain many recognition errors. In order to reduce the effect of topic detection errors a fallback mechanism is used where the initial hypothesis from the G-LM is compared with the topic-dependent decoding result. ASR-score is used to select the best hypothesis. 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, however topic detection accuracy declines. Training LMs for very narrow topics also generally suffer from data sparseness. Using wider topics increases topic detection accuracy, 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.

3. TOPIC DETECTION

In this paper, two topic detection methods are investigated. The first is based on unigram likelihood and the second based on SVM (Support Vector Machines) [6]. In both methods, models are created for each topic and topic detection takes place by comparing the sentence to be classified with each topic model and selecting the topic with the maximum score.

3.1 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 of the topic models against the 1-best hypothesis from the initial 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.

3.2 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 $(O(w_1), O(w_2),...,O(w_n))$, where $O(w_k)$ is the number of occurrences of word w_k in S_i. Feature vectors consist of 9600 features which 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. 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 comparing the vector representation of the sentence to be classified with 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. To provide improved robustness to ASR errors, the 10-best results from the initial recognition pass are used to produce the vector S_i. In this case, the value of $O(w_1)$ is the fraction of 10-best results that contain the word w_k.

4. TOPIC DEPENDENT LANGUAGE MODELING

The corpus used for evaluation had topic labels that were manually assigned. Using these labels to produce TD-LMs is not optimal in terms of either perplexity or topic detection accuracy. Thus automatic re-labeling methods are used to cluster the training data into topic dependent sets. The methods used for re-labeling are the same as those used for topic detection. In the case of unigram re-labeling, initial unigram models are created for each topic 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.

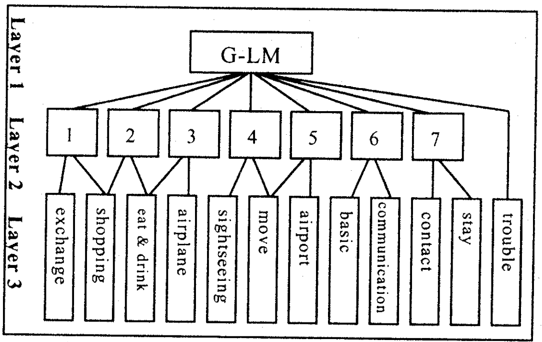


Figure 2: ***Multi-layer Language Modeling***

For SVM based re-labeling, this process is only done once. Initial topic models are created from the hand-labeled data and these models are used to automatically re-label the training set.

The re-labeling process reduces the LM perplexity of each topic by clustering similar sentences together. This in effect narrows the topic of each of the clusters and better models utterances within that topic. LMs are created for each topic, and 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 a development set.

5. MULTI-LAYER LANGUAGE MODELING

To increase the system's flexibility and robustness, a hierarchical LM framework is introduced. Intermediate language models are created to cover multiple topics. These topics can be detected more reliably than in the individual topic case, and are still expected to provide improved recognition compared to the G-LM. 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 hierarchy involves clustering together topics that are closely related. A distance measure relating to the topic detection method is used. These are described below.

5.2 Unigram based inter-topic distance

For unigram based topic detection, the distance between two topics is calculated as the normalized cross-perplexity as shown in Eq. (1).

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

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}

$DIST_{UNI}(C_i, C_j)$: Normalized cross perplexity of topics C_i, C_j

Here the perplexity of the training set of one topic is calculated in respect to the unigram model of the other, and vice-versa. These are normalized with respect to the original topic's perplexity. When the result is small, it indicates that the two topics are closely related.

5.3 SVM based inter-topic distance

For SVM based topic detection, the distance between two topics C_i, C_j is calculated as the average distance between one topic's training set and the other's SVM hyper-plane.

$$dist_{avg}(C_i, \vec{X}c_j) = \frac{1}{n_i}\sum_{k=0}^{n_i} dist(x_k, \vec{X}c_j) \qquad (2)$$

C_i : Topic class C_i

$\vec{X}c_i$: SVM hyper-plane for topic class C_i

$dist(x_k, \vec{X}c_i)$: perpendicular distance from SVM hyper-plane $\vec{X}c_i$ to point x_k, positive when in-class, and negative otherwise.

n_i : training set size of topic class C_i

$$DIST_{SVM}(C_i, C_j) = \left\| dist_{avg}(C_i, \vec{X}c_j) - dist_{avg}(C_j, \vec{X}c_j) \right\| + \left\| dist_{avg}(C_j, \vec{X}c_i) - dist_{avg}(C_i, \vec{X}c_i) \right\| \qquad (3)$$

The distance perpendicular to the SVM plane is used as this relates to the probability of occurrence of topic detection errors.

5.3. Multi-Layer Hierarchy Construction

The construction of the multi-layer hierarchy involves creating a set of intermediate nodes that cover multiple topics. Based on the appropriate distance measure hierarchical clustering involves finding the closest pair of topics and merging them.

The intermediate nodes created using this method relate to those topics that are most likely to be confused during topic detection. Moving up the hierarchy models cover an increasing number of topics, become less topic dependent, and are thus easier to detect, however the gain in recognition accuracy is also reduced.

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***

5.4. Topic Detection for Multi-layer Hierarchy

Within the topic detection stage we select an appropriate LM to use for re-decoding based on the recognition hypothesis from the initial recognition pass. If the result used for topic detection contains no recognition errors, then we can select the appropriate single topic TD-LM to re-perform decoding. However as recognition errors will occur, the correct topic cannot always be accurately selected. In these cases rather than selecting a detailed model that may cover an incorrect topic, it is more appropriate to select a model that is less topic dependent.

For unigram based topic detection, we create unigram models for each node in the hierarchy. Topic detection simply involves selecting the topic with the maximum unigram likelihood.

For SVM based topic detection, we select a layer 3 model when the SVM score for only one topic is positive otherwise we select the appropriate parent node of the best topic.

6. EXPERIMENTAL EVALUATION

The ATR phrasebook corpus [7] 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.

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 8.08%

6.1 Topic Dependent Language Modeling

Topic dependent language models (TD-LM) are created based on the original hand-labeled topic tags, these provide a 20.2% reduction in perplexity over a single G-LM (Table 2). This reduction verifies the effectiveness of topic dependent modeling. Next, unigram and SVM

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 (unigram labeled)	28.00 (37.5%)	16.85 (29.1%)
12 topics (SVM labeled)	29.6 (34.0%)	17.34 (27.1%)

Table 2: ***TD-LM perplexities***

Method	Layer 3	Any Layer
Unigram	90.2%	92.4%
SVM	92.3%	93.5%

Table 3: ***Topic detection performance***

	WER % (relative reduction)			
Classification Method	(G-LM) Layer 1	Layer 3	Layer 1,3	All Layers
Unigram based Topic Detection				
Topic Known (Oracle)	8.08	7.36 (8.9%)	6.33 (21.7%)	6.30 (21.8%)
ASR Based Topic Detection	8.08	8.12 (-0.5%)	7.36 (8.9%)	7.30 (9.7%)
SVM based Topic Detection				
Topic Known (Oracle)	8.08	7.64 (5.2%)	7.10 (12.0%)	7.08 (12.0%)
ASR Based Topic Detection	8.08	8.24 (-1.2%)	7.42 (8.2%)	7.25 (10.3%)

Table 4: ***System recognition performance***

re-labeling 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 which takes into account the size of the topic clusters, and tends to balance the training set evenly over the 12 topics. The smallest topic is 35% the size of the largest topic. For the SVM based method, topics are clustered without consideration to cluster size. In this case, the smallest topic is only 8% that of the largest. It is also confirmed that in the SVM case the resulting topics are related better to the original topic labels.

6.2. Topic Detection Accuracy

Next, we investigate the detection performance of the topics defined in the previous process. The topic detection accuracy is evaluated by comparing the ASR based topic detection result with that based on the original transcription which is 100% correct. Table 3 shows the topic detection accuracies of the unigram and SVM based methods. The first column gives the accuracy of selecting a single layer 3 topic, and the second column gives the topic detection accuracy when a multi-layer hierarchy is used. In this case, the accuracy is that of selecting either a layer 3 topic, or a layer 2 model that includes that topic. SVM based topic detection provides increased robustness in the face of recognition errors when compared with the unigram method. When a multi-layer hierarchy is used as described in section 5 the topic detection accuracy increases by around 2%.

6.3. Unigram based Topic Dependent Recognition

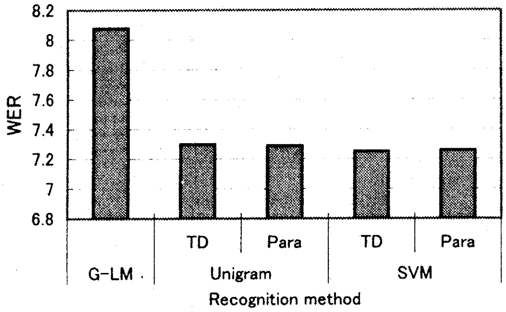
Next, the system performance using the unigram based topic detection methods is investigated. Recognition is performed in two stages as described in section 2. The performance when using different topic layers is given in Table 4 (upper half). As the G-LM model is always applied in the initial recognition pass, 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.

In the case of oracle topic detection, the most detailed layer 3 models provide a gain of 8.9% over the baseline system. By including the comparison with the layer 1 G-LM model, this gain is increased to 21.7%. 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 is trained over the entire training set, and is thus less affected by data sparseness than the individual topic models. When using an oracle for topic detection, the inclusion of the intermediate layer 2 models provides little gain in recognition accuracy.

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 system. Introducing the comparison with layer-1 (G-LM) mitigates the effect of topic detection errors, and the WER is reduced by 8.9% relatively. With the inclusion of the layer 2 models, the accuracy of selecting a topic or its parent is 92.4%. This method provides a 9.7% relative reduction in WER over the baseline system.

6.3. SVM based Topic Dependent recognition

The system performance using SVM based topic detection is given in Table 4 (lower half). When the topic is given by an oracle, the system performance is much lower than the unigram method. However for ASR based topic detection, when layers 1 and 3 are used for recognition, a



G-LM: Single Generalized LM (baseline)
TD: Topic dependent model selection using ASR based topic detection
PARA: Use G-LM + all 12 TD-LM in parallel

Figure 5: ***Comparison of ASR performance***

reduction in WER of 8.2% is gained, and the inclusion of the layer 2 models increases this reduction to 10.3%. For both topic detection methods the proposed architecture provides a relative reduction in WER of around 10% over the baseline system.

6.4. Parallel System Performance

Finally, the proposed system is compared with a parallel system where recognition is performed with all the individual TD-LMs and the G-LM in parallel. The hypothesis 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.

The baseline system has a WER of 8.08%. Independent of the topic detection method used both the proposed and parallel systems reduce the WER to between 7.25% and 7.30%, a 10% relative reduction over the baseline system. There is no significant difference between the WER of these methods. While the proposed and parallel systems achieve comparable recognition performance, the computational cost of the proposed method is 1/6th of that of the parallel system.

7. DISCUSSION

When an oracle is used for topic detection, the unigram method significantly outperforms SVM. As unigram labeling creates topics of similar size, it appears to create better trained models than in the SVM based method. However, when using ASR based topic detection, the SVM based method provides increased topic detection accuracy and the performance of the proposed architecture is similar for the two methods.

The multi-layer hierarchy introduced in this paper provides increased topic detection accuracy, and a small reduction in WER. Independent of the topic detection method, the proposed system provides similar performance to a parallel system, while requiring only 1/6th of the computational cost.

8. CONCLUSIONS

We have presented an efficient speech recognition architecture based on topic detection and topic-dependent language modeling. The proposed system provides up to a 10.3% relative reduction in WER over a single LM system and achieves recognition performance that is comparable to running a large number of LMs in parallel while requiring a much smaller computational cost. Two topic detection methods, unigram likelihood, and SVM are compared, and both methods provide a similar improvement in recognition accuracy. A novel framework of multi-layer LMs is also introduced. This framework provides increased robustness against topic 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 dialog translation technology based on a large corpus." The authors are grateful to Prof. H. G. Okuno of Kyoto Univ. for his useful comments.

9. 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] S. C. Martin et al, "Adaptive Topic-Dep. Language Modelling Using word based Varigrams", Proc. EUROSPEECH '97, Vol. 3, pp. 1447-1450, 1997

[6] T. Joachims, "Text Categorization with Support Vector Machines", Proceedings of the European Conference on Machine Learning, Springer, 1998

[7] 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