ROBUST TOPIC INFERENCE FOR LATENT SEMANTIC LANGUAGE MODEL ADAPTATION

Aaron Heidel and Lin-shan Lee

Dept. of Computer Science & Information Engineering
National Taiwan University
Taipei, Taiwan, Republic of China

aaron@speech.ee.ntu.edu.tw lslee@gate.sinica.edu.tw

ABSTRACT

We perform topic-based, unsupervised language model adaptation under an N-best rescoring framework by using previous-pass system hypotheses to infer a topic mixture which is used to select topic-dependent LMs for interpolation with a topic-independent LM. Our primary focus is on techniques for improving the robustness of topic inference for a given utterance with respect to recognition errors, including the use of ASR confidence and contextual information from surrounding utterances. We describe a novel application of metadata-based pseudo-story segmentation to language model adaptation, and present good improvements to character error rate on multi-genre GALE Project data in Mandarin Chinese.

Index Terms— language model adaptation, topic modeling, unsupervised adaptation, speech recognition, story segmentation

1. INTRODUCTION

For over 20 years, statistical n-gram-based language models have been an effective way to model human language for tasks in both speech-to-text (STT) and information retrieval (IR) applications. Even so, the technique has its weaknesses, most notable of which is an inability to handle long-range context. Essentially, each language model is trained for a single domain; hence if the test data comes from multiple domains, the best such language model we can come up with is a "jack of all trades, master of none" language model. That is, a language model that performs tolerably for everything, and excellently for nothing.

Fortunately, word co-location-based techniques such as Latent Semantic Analysis (LSA) and its derivatives Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) offer us the tools to explicitly model coarse-grain language context, or topics [1] [2] [3]. In consequence we have a well-founded mechanism for performing *language model adaptation*, where we somehow guess the topic — or some mixture thereof — of a piece of text and use that knowledge to adjust the language model to fit.

We propose an unsupervised topic-based language model adaptation scheme that extends and improves on previous work [4] by making run-time topic inference more robust to recognition errors. Using a topic model we perform an utterance-level decomposition of a heterogeneous text training corpus into many topic-specific text corpora, each of which is used to estimate a corresponding topic-specific n-gram language model. We demonstrate ways to segment a sequence of consecutive utterances into *topical context windows* and use these windows to recover from the recognition errors in system hypotheses when performing topic inference on each consistuent utterance. The inferred topic mixture is then used to select a set of relevant topic LMs for interpolation with a topic-independent background language model and also to set the interpolation weight for each topic LM. We perform language model adaptation under an N-best rescoring framework.

2. RELATED WORK

One of the earliest attempts to perform language model adaptation was the cache-based technique, which boosts the probabilities of words recently observed [5]. This technique was then generalized using trigger pairs, in which the observation of certain "trigger" words increases the probability of seeing correlated words [6].

Another well-known approach is the sentence-level mixture model, which used topics identified from a heterogenous training corpus by automatic clustering [7]. Improvements were demonstrated in both perplexity and recognition accuracy over an unadapted trigram language model.

The story topic-based approach to large-scale, fine-tuned language model adaptation [8] is similar to ours in the construction of a set of topic LMs from a heterogeneous training corpus and their linear interpolation with a background language model. Their approach differs from the proposed approach in three primary ways: (1) manually-defined article keywords are taken as topic labels; (2) TF-IDF and naive Bayes classifiers are used for topic inference; and (3) a large (5000) number of topics are defined, while the experiments reported here use 64.

The approach described in [9] uses both a mixture-based

model and an exponentially decaying cache to adapt a trigram language model. Our approach could be seen as a refinement of the general ML-based mixture model using the MAP-based Latent Dirichlet Allocation (LDA) topic model, with special emphasis placed on robust topic inference given unreliable data.

A state-of-the-art approach to language model adaptation is [10] [11], where the background language model is adjusted to fit a set of LDA or Latent Dirichlet-Tree Allocation (LDTA) based marginals. This work boasts an elegant formulation and appears to be very efficient; we believe there could be room for even greater improvement by adapting the background language model according to n-gram-based — as opposed to unigram-based — constraints.

Another recent approach is described in [12], where they report on techniques for unsupervised language model adaptation for the broadcast conversation transcription task. They investigate the effect when small amounts of in-domain (i.e., broadcast conversation) data are added to a large, general-domain (i.e., broadcast news) LM training corpus, and perform a valuable comparison of PLSA- and LDA-based LM adaptation, concluding that there is little difference between the two methods in terms of character error rate.

In contrast to the above two approaches, the proposed method decomposes the background LM into topic LMs using utterance-level n-gram counts. As such, the proposed method is different from all such approaches that directly manipulate the background LM according to some unigram distribution based on the adaptation text.

This approach is also conceptually simpler than a recent work on language model adaptation for lectures using HMM-LDA [13], for example, in that no distinction is made between syntactic and semantic states.

3. METHODOLOGY

The proposed approach has four main components: two performed off-line and two on-line. The off-line components are topic model training and topic language model estimation, while the two on-line components are topic inference and language model interpolation. Also, it should be noted that while our experiments were all performed using the LDA topic model, the approach is in fact independent of the topic model type used.

3.1. Topic Model Training

Latent Dirichlet Allocation is a generative, probabilistic model characterized by the two sets of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\alpha} = [\alpha_1 \alpha_2 \cdots \alpha_k]$ represents the Dirichlet parameters for the k latent topics of the model, and $\boldsymbol{\beta}$ is a $k \times V$ matrix where each entry β_{ij} represents the unigram probability of the jth word in the V-word vocabulary under the ith latent topic. As described in [4], the iterative LDA topic inference algorithm takes as input a bag (or set) of words $\boldsymbol{w}$ and an initial topic mixture $\boldsymbol{\theta}^{[0]}$ and returns a vector $\boldsymbol{\theta} = [\theta_1 \theta_2 \cdots \theta_k]$ containing the topic mixture weights. The initial topic mixture $\boldsymbol{\theta}^{[0]}$ corresponds to the topic distribution of the topic model's original training corpus.

Since LDA is a supervised model, and we are not generally supplied with labeled training corpora, we construct one in an unsupervised manner using PLSA. We then train the LDA model using the PLSA-derived topic-document mappings as an initial model.

3.2. Topic Language Model Estimation

After training our topic model, we proceed to classify each individual utterance in the training corpus as belonging to one of the k topic corpora as follows: for each such utterance, we infer the topic mixture $\boldsymbol{\theta}$ from which we choose the topic with the maximum weight, and append the utterance to this topic's corpus. We then use the resulting k topic-specific corpora to train each topic LM[1]. In our experiments, the SRILM toolkit was used for all language model training and interpolation [14].

3.3. Robust Topic Inference

As in many other unsupervised language model adaptation schemes, we use previous-pass system hypotheses as our adaptation text from which we determine in what "direction" the language model should be adapted. Various previous works have demonstrated the problem of erroneous hypotheses: the errors we seek to recover from often lead our adapted language models astray and thus result in severely degraded performance. Hence the primary focus of this work is how to compensate for this. How do we improve the robustness of unsupervised language model adaptation — in our case, how do we design our topic inference mechanism to keep the good and throw out the bad? Or, equivalently, how do we pull the LM only toward those topics represented by the parts of the hypothesis that we are confident about, and not toward the topics represented by the parts we are unsure of?

When using the LDA topic model, a straightforward approach to improving the robustness of topic inference is to alter the topic inference algorithm (1) to allow for arbitrary initial topic mixtures and (2) to allow for bags of arbitrarily-weighted words, as opposed to the conventional bags of uniformly-weighted words.

3.3.1. Custom Prior Mixtures

When we want to infer a topic mixture for a word sequence $\boldsymbol{w}$, we start from the initial topic mixture $\boldsymbol{\theta}^{[0]}$ and iteratively adjust the mixture according to the words in $\boldsymbol{w}$ until the mixture converges [4]. In some cases, we believe that this initial mixture may be unnecessarily broad. For instance, if we want to infer the topic mixture for the utterance u, and we happen

[1]Note the distinction between topic LMs, which are highly targeted toward a single topic, and the background LM, which is a general-domain, topic-independent LM.

to know that the set of utterances in question is mostly about sports, it makes sense to bias the initial mixture accordingly.

In an unsupervised framework, this could be as simple as inferring a topic mixture $\hat{\boldsymbol{\theta}}_u$ for the utterances surrounding the utterance u — its *topical context window* — and using that topic mixture as an initial topic mixture — or a *prior* topic mixture, in Bayesian parlance — for the inference algorithm when inferring u's topic mixture. This corresponds to replacing the LDA model's $\boldsymbol{\alpha}$ vector with one of our own choosing by replacing $\boldsymbol{\theta}^{[0]} = \left[\frac{\alpha_1}{\alpha_{sum}} \frac{\alpha_2}{\alpha_{sum}} \cdots \frac{\alpha_k}{\alpha_{sum}}\right]$ with $\boldsymbol{\theta}^{[0]} = \hat{\boldsymbol{\theta}}_u$.

The resulting challenge becomes defining the topical context window: smaller such windows tend to reinforce the recognition errors we seek to recover from, while larger windows take us back to the original problem of unnecessarily broad initial mixtures.

3.3.2. Topical Context Window Segmentation

Optimal topical context window segmentation results in reasonable prior mixtures for every consistuent utterance; as such, these segments should match the division of stories, or topics, within the input set. There are two types of segmentation: content-based and metadata-based, where *content* refers to the contents of a given utterance and *metadata* refers to information about the utterance. In this work we do not report on content-based schemes.

In our experiments, since the corresponding filename for the N-best list of each utterance contained useful metadata, we used metadata-based segmentation. The filename contains the elements `PROGRAM`, `START`, and `END`, where `PROGRAM` refers to program names such as "CCTV4 DAILYNEWS 2006/11/13", and `START` and `END` refer to the start and end timestamps of the utterance in question. We define the timegap between two consecutive utterances u_i and u_{i+1} as the difference between the end timestamp of u_i and the start timestamp of u_{i+1}. Clearly, segments should be broken at explicit program breaks; we may additionally break segments when the timegap is greater than a threshold, because this may indicate deleted commercials or the like that could indicate a change of context.

3.3.3. Custom Word Weights

Under the LDA model, we assume that each word is an equally reliable observation and thus the posterior probability of each word in $\boldsymbol{\beta}$ has equal weight in determining the topic mixture. However, this assumption does not hold when using erroneous data for inference. By relaxing the assumption of uniform reliability, we can replace $\boldsymbol{\theta}_i^{[t+1]} = \frac{1}{M} \sum_{n=1}^{M} \kappa_{ni}$ with

$$\boldsymbol{\theta}_i^{[t+1]} = \frac{1}{\sum_{j=1}^{M} \gamma_j} \sum_{n=1}^{M} \gamma_n \kappa_{ni},$$

where γ_n is the weight for word w_n. With this alteration, we can use whatever confidence information we have from the previous-pass recognition system as an estimate of the relative reliability of each word in $\boldsymbol{w}$.

3.3.4. Weighted N-Best Topic Inference

In particular, for our experiments, the N-best lists outputted from the recognition system contained acoustic model and language model scores AM_l and LM_l for each hypothesis l. We estimate the confidence value γ_n for each distinct word w_n appearing in the N-best list in the following way: for each occurrence of w_n in the N-best list, we add $post_l^{\frac{conf}{wc_l}}$ to γ_n, where $post_l$ and wc_l are the posterior probability and word count for hypothesis l. $conf$ is the confidence weight for N-best topic inference: at $conf = 0$ words are weighted only by their frequency in the N-best list and at $conf = 1$ they are weighted according to posterior probability. The posterior probability is computed from the acoustic and language model scores as $post_l = AM_l + lmw * LM_l$. In the reported experiments, the language model weight *lmw* was set to 6.5.

3.3.5. Utterance Decay

When we seek to infer the topic mixture for a given utterance, we generally consider only the words in the given utterance. However, it makes sense to also consider the words in surrounding utterances (provided they are within the topical context window), because words mis-recognized for the given utterance may be recognized correctly in surrounding utterances; doing so more closely reflects the way humans use context to recover from recognition errors. Moreover, we can weight these surrounding words by a decay factor, presumably such that the words from the given utterance are assigned the greatest significance.

Thus we use utterance decay to specify the weights given to words in surrounding utterances. Where u_i is the utterance whose topic mixture we wish to infer, the weight for each of the words in utterance u_j when inferring $\boldsymbol{\theta}_{u_i}$ is set to $decay^{|i-j|}$. Note that these weights are independent of the confidence weights described in Section 3.3.4.

Hence a decay of 0 would correspond to the approach described in our previous work, where surrounding utterances are totally ignored, while a decay of 1 would correspond to the same topic mixture being inferred for each utterance in the topical context window, which would mean the same adapted LM is used for all utterances in a given topical context window.

3.4. Language Model Interpolation

Once we have determined the topic mixture for a given utterance u, we can proceed to adapt the language model and use it to rescore the corresponding N-best list. Here we use the topic mix threshold weight *tmtw* parameter as a threshold for assembling a set of relevant topic LMs. Thus, where λ_B is the (static) interpolation weight for the background LM, the interpolation weight λ_i for topic i's LM is set to

$$\lambda_i = (1 - \lambda_B) \frac{\lambda_i'}{\sum_{j=1}^{k} \lambda_j'}, \text{ where } \lambda_i' = \begin{cases} 0 & \text{if } \boldsymbol{\theta}_i < tmtw \\ \boldsymbol{\theta}_i & \text{otherwise.} \end{cases}$$

Source Name	Stories	Utterances
GALE	42,265	858,871
Chnews	—	8225
Downloaded-Web-Data	362,630	5,964,747
Giga_2005T14-cna	—	17,868,725
Giga_2005T14-xin	—	13,815,340
Giga_2005T14-zbn	—	802,485
MTC123	—	2528
TDT[234]	2585	733,738
TOTAL	407,480	40,054,659

Table 1. Broadcast news (BN) training data. Stories are those explicitly defined in training data.

Source Name	Stories	Utterances
Downloaded-Web-Data	2302	90,797
GALE	524	683,192
TOTAL	2826	773,989

Table 2. Broadcast conversation (BC) training data.

	Unigrams	Bigrams	Trigrams	4-grams	TOTAL
full	60421	58 M	316 M	201 M	575 M
pruned	60421	19.4 M	24.2 M	6.1 M	49.8 M

Table 3. 4-gram background LM n-gram counts.

4. EXPERIMENTAL SETUP

All of the reported experiments were performed on data as part of Phase II of DARPA's GALE program[2]. We evaluated the proposed LM adaptation approach on NIGHTINGALE [15] — the UW-SRI-ICSI Mandarin broadcast speech recognition system — under an N-best rescoring framework. We evaluated the approach using the `dev07`[3] development set, which contains 1736 utterances and is composed of approximately 60% broadcast conversation (BC) and 40% broadcast news (BN) genres. We used the `dev07a` subset (containing 719 utterances) of `dev07` for parameter tuning and evaluated on the entire `dev07` set. Tables 1 and 2 list all of the text training data provided by the Linguistic Data Consortium (LDC) for the GALE Program, used for topic model and language model training.

The 4-gram background LM, part of NIGHTINGALE, was trained using the modified Kneser-Neys smoothing scheme [16]. Due to memory constraints, we used a pruned version of this model in experiments instead of the full 4-gram model. Table 3 shows the number of explicit n-gram parameters before and after pruning the model using a 10^{-9} entropy threshold.

[2]See `http://www.darpa.mil/ipto/programs/gale/` and `http://projects.ldc.upenn.edu/gale/`.

[3]The IBM-modified version, not the original LDC version.

Note again that this language model is a topic-independent, general-domain language model that we interpolate with a set of topic-dependent language models when performing LM adaptation.

Some of the following experiments are performed for supervised language model adaptation, in which the reference transcripts are used for topic inference. Thus supervised experiments represent — at least in the sense of topic inference — the upper bound for the performance of the proposed approach, where the LM is biased toward the correct answer, or "oracle".

5. RESULTS

5.1. Topic Model Training

From the training data, we extracted a set of 64,029 topic-coherent documents, or stories, for use in training the topic model. Of the 407 K explicitly-marked BN stories, we randomly selected 51 K for topic model training. For the BC data, since the explicitly-marked stories were both few in number and long in length, we broke many of these longer stories into several smaller stories, resulting in 5.5 K stories, and also broke a remaining 366 K BC-genre utterances into 7.3 K 50-utterance pseudo-stories. Thus, we used 64 K stories total for topic model training: 51,223 BN- and 12,806 BC-genre stories. This resulting 4:1 BN-to-BC ratio is in notable contrast to the 50:1 ratio observable in the entire training set; this was to allow for better detection of BC topics. The vocabulary size was approximately 60 K.

As described in Section 3.1, we trained a 64-topic, 20-iteration PLSA model as an initial model for the LDA topic model. Table 4 lists a few representative topic descriptions from the final LDA topic model (the topic descriptions were generated using word entropy over all topics multiplied by word posterior probability). This selection of topics is sorted by decreasing frequency in the entire training data. Here topic 36 stands out as the most frequent topic: clearly this is due to the `cna` Gigaword corpus from Taiwan, the largest corpus in our training data. Topic 19, on the other hand, seems to come from celebrity interview transcripts on programs like Phoenix TV's 鲁豫有约 ("Date With Lu-Yu"). The top 4 topics represent the BN genre while the topics 19 and 3 correspond to the BC genre. Of the 64 topics, 46 and 18 topics could be considered BN- and BC-genre topics, respectively.

We trained 64 4-gram topic language models using modified Kneser-Neys backoff using the procedure described in Section 3.2.

5.2. Timegap Threshold

Figures 1 and 2 show the effects of different timegap thresholds on character error rate (CER). Note that average segment size increases with the timegap threshold. Figure 1 shows results for unsupervised adaptation using frequency-weighted N-best topic inference for utterance decay at 0.65 and 1: here we observe that setting decay to 1 makes for consistently poorer performance than that for $decay = 0.65$. In addition,

ID	Most Significant Words in Topic
36	民进党 陈水扁 国民党 马英九 台湾 DPP, Chen Shui-bian, KMT, Ma Ying-jeou, Taiwan
54	增长 美元 价格 油价 石油 原油 市场 increase, USD, prices, oil price, oil, crude oil, market
4	比赛 赛 队 杆 冠军 选手 球 match, compete, team, club/stick, champion, athlete, ball
32	香港 曾荫权 立法会 行政长官 Hong Kong, Donald Tsang, LegCo, Commissioner
19	她 我 拍 周杰 歌迷 戏 演 she, I, photographed, Zhou Jie, fans, drama, perform
3	我 你 就 这个 啊 她 他 那个 那 了 呢 去 着 吃 钱 没 I, you, then, this, she, he, that, so, went, ate, money, no

Table 4. Several topic descriptions.

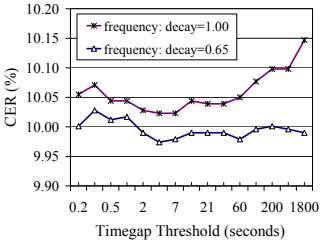

Fig. 1. Unsupervised adaptation given timegap threshold. $\lambda_B = 0.5$.

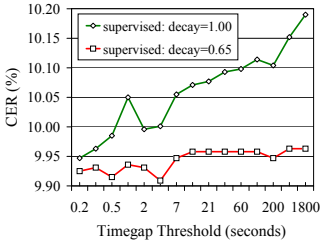

Fig. 2. Supervised adaptation given timegap threshold. $\lambda_B = 0.1$.

larger segments result in significantly degraded performance for $decay = 1$ but have less of an impact for $decay = 0.65$. In contrast, smaller segments lead to degraded performance in both cases. We observe a general dip toward an optimal threshold around 4 seconds for both curves.

For supervised adaptation, as shown in Figure 2, in general, the smaller the segment the better.

5.3. Topic Inference

Figure 3 shows the effect of utterance decay separated for four different types of topic inference: oracle-based, inference based on the top single system hypothesis ("1-best"), and frequency- and posterior-based N-best topic inference. Here we see that the performance of unsupervised adaptation improves as utterance decay increases, but that of supervised adaptation degrades. These results are similar to those for Figure 2, and make sense, as supervised adaptation does not have to deal with recognition errors and should thus achieve theoretically perfect topic inference at $decay = 0$; higher $decay$ values only tend to confuse topic inference.

As can be seen in Figures 3 and 4, frequency-based N-best topic inference consistently outperforms its posterior-based alternative.

Figures 3 and 4 also show results when basing topic inference on only the top-1 system hypothesis as compared to that using N-best-based topic inference. As would be expected,

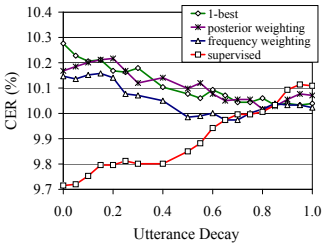

Fig. 3. Unsupervised vs. supervised adaptation given utterance decay. λ_B set to 0.5 (0.1) for unsupervised (supervised) experiments.

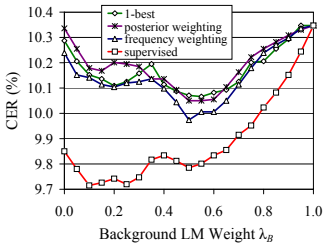

Fig. 4. Unsupervised vs. supervised adaptation given λ_B. Utterance decay set to 0.65 (0) for unsupervised (supervised) experiments.

LM	PLP	ICSI
full 3-gram	12.0%	11.9%
full 4-gram	11.9%	11.7%
adapted 4-gram	11.7%	11.4%

Table 5. Final results for `dev07`.

it pays to take into account the complete N-best list when performing topic inference.

The results for experiments on custom prior mixtures are not shown here, as their effect was inconsistent and insignificant. Utterance decay and the background LM weight λ_B influence CER far more than the choice of prior mixture.

As seen in Figure 4, the most important parameter for this scheme is the background LM weight λ_B, which represents how much our adapted LM depends on the background LM.

5.4. Final Results

Table 5 shows the final CER results for `dev07` on NIGHTINGALE, which is composed of two recognition systems (PLP and ICSI) with different error patterns for use in system combination.

Here we see that unaltered (that is, generated with a full topic-independent trigram LM), the N-bests have CERs of 12.0% and 11.9% for PLP and ICSI, respectively. When we perform N-best rescoring with the static (no LM adaptation), unpruned 4-gram LM, we obtain CERs of 11.9% and 11.7%. However, when we perform rescoring using adapted LMs ($\lambda_B = 0.5$, $decay = 0.8$, with custom prior topic mixtures), we obtain CERs of 11.7% and 11.4%. Note that this is not only better results than that using the full static 4-gram LM, but it also comes at a much lower price in terms of memory and CPU. Specifically, the 580 M parameters of the full 4-gram background LM require more than 8 GB of memory, while our adapted LM — which contains less than a tenth the number of parameters — requires less than 700 MB of memory, and runs at the rate of approximately 0.4 × RT on a single 3 GHz CPU core. Thus the proposed approach clearly succeeds in "getting more bang for the buck" in biasing the LM toward what is reasonable, given previous-pass system hypotheses.

6. DISCUSSION

Utterance decay is shown to be highly effective in recovery from topic inference bias caused by recognition errors by widening the net to allow for better topic inference, in the sense that a single utterance's idiosyncracies — or recognition errors — have less of an influence on the resultant topic mixture. Decay also seems to be closely linked to the segmentation of topical context windows. That is, the closer utterance decay is to 1, the more we rely on topical context window segmentation to limit the contents of our weighted bag of words $\boldsymbol{w}$ to those words that are really contextually relevant; in contrast, the closer decay is to 0, the less of a role such segmentation plays. Thus future work will include the investigation of content-based segmentation for applications where metadata is not available, and also principled ways to integrate metadata- and content-based segmentation.

It is not known why, for N-best-based confidence measures, frequency-based confidence outperforms posterior-based confidence. This issue deserves further investigation. In addition, custom prior mixtures for topic inference were found to be of inconsistent utility.

We believe that the proposed approach is conceptually sound and constitutes a simple but effective approximation of human cognitive processes when performing speech recognition. Among N-best, word graph, and confusion network rescoring, it is reasonable that N-best rescoring affords the smallest improvements; thus future work will include rescoring for richer search-space representations.

In general, these results show the dependence of the approach on proper tuning. On one hand, this is to be expected, considering the higher semantic level of information we are dealing with. On the other hand, it would be desirable to find ways to base these parameters as much as possible on the content itself and not exclusively on development sets. This is another direction for future work.

7. CONCLUSION

We have described improvements to earlier work on unsupervised topic-based LM adaptation that render such adaptation less susceptible to the misleading effects of previous-pass recognition errors, including the judicious use of ASR confidences and contextual information via utterance decay, which together serve to constrain inferred topic mixtures to what is reasonable. We also introduced a useful application of pseudo-story segmentation in defining topical context windows for the LM adaptation task.

Good improvements to character error rate were demonstrated for the challenging multi-genre (BN/BC) speech-to-text task, despite the limited N-best rescoring framework. Great potential for further improvements exists for future work using richer search-space representations such as word graphs and confusion networks, as well as when combined with the use of more sophisticated techniques for topical context window segmentation.

It is not known to what extent the proposed approach depends on the type of topic model used. Thus, future work may include experiments to see what advantages the MAP-based LDA model really brings as opposed to the ML-based PLSA or the classical mixture model described in [9]. Would LDA-based LM interpolation weight determination really result in better performance than simpler EM-based alternatives?

8. REFERENCES

[1] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society of Information Science*, 1990.

[2] T. Hofmann, "Probabilistic Latent Semantic Analysis," *Uncertainty in Artificial Intelligence*, 1999.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, 2003.

[4] A. Heidel, H. A. Chang, and L. S. Lee, "Language Model Adaptation Using Latent Dirichlet Allocation for Topic Inference," in *Proceedings of Interspeech*, 2007.

[5] R. Kuhn and R. De Mori, "A Cache-Based Natural Language Model from Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990.

[6] R. Rosenfeld, "A Maximum Entropy Approach to Adaptive Statistical Language Modeling," *Computer, Speech and Language*, 1996.

[7] R. Iyer and M. Ostendorf, "Modeling Long Distance Dependency in Language: Topic Mixtures vs. Dynamic Cache Models," in *Proceedings of ICSLP*, 1996.

[8] K. Seymore and R. Rosenfeld, "Using Story Topics for Language Model Adaptation," in *Proceedings of Eurospeech*, 1997.

[9] P. R. Clarkson and A. J. Robinson, "Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache," in *Proceedings of ICASSP*, 1997.

[10] Y. C. Tam and T. Shultz, "Unsupervised Language Model Adaptation Using Latent Semantic Marginals," in *Proceedings of Interspeech*, 2006.

[11] Y. C. Tam and T. Shultz, "Correlated Latent Semantic Model for Unsupervised LM Adaptation," in *Proceedings of ICASSP*, 2007.

[12] D. Mrva and P. C. Woodland, "Unsupervised Language Model Adaptation for Mandarin Broadcast Conversation Transcription," in *Proceedings of ICSLP*, 2006.

[13] B. J. Hsu and J. Glass, "Style & Topic Language Model Adaptation Using HMM-LDA," in *EMNLP*, 2006.

[14] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proceedings of ICSLP*, 2002.

[15] M. Y. Hwang, G. Peng, W. Wang, A. Faria, and A. Heidel, "Building a Highly Accurate Mandarin Speech Recognizer," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2007.

[16] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," in *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, 1996.