UNSUPERVISED LANGUAGE MODEL ADAPTATION VIA TOPIC MODELING BASED ON NAMED ENTITY HYPOTHESES

Yang Liu, Feifan Liu

The University of Texas at Dallas, Richardson, TX, USA

{yangl,ffliu}@hlt.utdallas.edu

ABSTRACT

Language model (LM) adaptation is often achieved by combining a generic LM with a topic-specific model that is more relevant to the target document. Unlike previous work on unsupervised LM adaptation, in this paper we propose to leverage named entity (NE) information for topic analysis and LM adaptation. We investigate two topic modeling approaches, latent Dirichlet allocation (LDA) and clustering, and proposed a new mixture topic model for LDA based LM adaptation. Our experiments for N-best list rescoring have shown that this new adaptation framework using NE information and topic analysis outperforms the baseline generic N-gram LM based on a state-of-the-art Mandarin recognition system.

Index Terms— language model adaptation, latent Dirichlet allocation (LDA), clustering, named entities, rescoring.

1. INTRODUCTION

Language model (LM) adaptation is important in speech recognition in order to better deal with a variety of topics and styles. When topic information is unavailable, an unsupervised LM adaptation approach typically first performs latent topic analysis, and combines the topic specific model with the generic N-gram. To identify implicit topics from an unlabeled corpus, one simple technique is to group the documents into topic clusters by assigning only one topic label to a document [1]. Recently several other methods in the line of latent semantic analysis have been proposed and used in LM adaptation, such as latent semantic analysis (LSA) [2], probabilistic latent semantic analysis (PLSA) [3], and LDA [4]. Most of these existing approaches are based on the "bag of words" model to represent documents, where all the words are treated equally and no relation or association between words is considered.

Unlike prior work in LM adaptation, this paper investigates how to effectively leverage named entity information for latent topic analysis in speech recognition. Named entities are very common in domains such as broadcast news, and carry valuable information, which we hypothesize is topic indicative and useful for latent topic analysis. We investigate two topic modeling approaches for LM adaptation, LDA and clustering, both using named entity information. For LDA, we also propose a topic mixture model based on the word, document, and topic distribution. In our prior work [5], we only evaluated the proposed unsupervised LM adaptation method using perplexity. Since perplexity does not always correlate well with recognition performance, in this study, we evaluate whether this proposed approach improves speech recognition on a large vocabulary continuous speech recognition task using a state-of-the-art recognizer. Our experiments show that using the new LM adaptation for N-best list rescoring achieves slightly better recognition performance than a 3-gram or 5-gram LM, trained from about 700 million words.

The rest of this paper is organized as follows. In Section 2 we review some related work. Section 3 describes in detail our unsupervised LM adaptation approach using named entities. N-best list rescoring results are presented and discussed in Section 4. Conclusion and future work appear in Section 5.

2. RELATED WORK

For unsupervised LM adaptation, an early attempt is a cache-based model [6], developed based on the assumption that words appearing earlier in a document are likely to appear again. The cache concept has also been used to increase the probability of unseen but topically related words, for example, the trigger-based LM adaptation using the maximum entropy approach [7].

Latent topic analysis has recently been investigated extensively for language modeling. Iyer and Ostendorf [1] used hard clustering to obtain topic clusters for LM adaptation, where a single topic is assigned to each document. Bellegarda [2] employed latent semantic analysis (LSA) to map documents into implicit topic sub-spaces and demonstrated significant reduction in perplexity and word error rate (WER). Its probabilistic extension, PLSA, is powerful for characterizing topics and documents in a probabilistic space and has been used in LM adaptation [3]. Proposed by Blei et al. [4], latent Dirichlet allocation (LDA) loosens the constraint of the document-specific fixed weights by using a prior distribution and has quickly become one of the most popular probabilistic text modeling techniques. LDA has been shown to outperform PLSA in corpus perplexity and text classification experiments (e.g., [4]). Tam and Schultz [8] successfully applied the LDA model to unsupervised LM adaptation by interpolating the background LM with the dynamic unigram LM estimated by the LDA model. Hsu and Glass [9] investigated using hidden Markov model with LDA to allow for both topic and style adaptation. Mrva and Woodland [10] achieved a WER reduction on broadcast conversation recognition using an LDA based adaptation approach that effectively combined the LMs trained from corpora with different styles: broadcast news and broadcast conversation data. Heidel et al. [11] used an efficient topic inference algorithm for LDA and achieved lower perplexity and improved recognition accuracy.

The focus of our work is to investigate the role of named entity information for topic modeling and LM adaptation. This is different from using all the words or selecting terms for topic analysis as those used in text categorization or information retrieval. In [5], we investigated unsupervised LM adaptation using clustering and LDA based topic analysis approaches. We also proposed a novel

dynamic weighting scheme in the LDA based framework for topic adapted LM, different from [8]. However, experiments were only conducted to measure the perplexity. In this study, our goal is to examine whether the gain we observed for the perplexity also extends to recognition results.

3. UNSUPERVISED LM ADAPTATION USING NAMED ENTITIES

We evaluate two different topic modeling approaches for LM adaptation, LDA and clustering, both using NE hypotheses. The following sections explain in detail the training and testing procedure.

3.1. Training

For training, we use the text collection to train the generic word-based N-gram LM. Each document in the training set is labeled with NE hypotheses using a named entity tagger. Then we perform topic analysis using these NEs and train multiple topic specific N-gram LMs.

3.1.1. LDA

The purpose of LDA analysis in training is to find the latent topic information for the given document collection. We use the MATLAB topic Toolbox 1.3 [12] on the training set to obtain the document-topic matrix, DP, and the word-topic matrix, WP. Note that here "words" correspond to the elements used to represent the document (i.e., NEs in our experiments). In the DP matrix, an entry c_{ik} represents the counts of words in a document d_i that are from a topic z_k ($k = 1, 2, \ldots, K$). In the WP matrix, an entry f_{jk} represents the frequency of a word w_j generated from a topic z_k ($k = 1, 2, \ldots, K$) over the training set.

After LDA analysis, we use a hard decision to create topic clusters by assigning a topic z_i^* to a document d_i such that

$$z_i^* = \underset{1 \le k \le K}{\operatorname{argmax}}\, c_{ik}.$$

Based on the documents belonging to each topic cluster, K topic N-gram LMs are trained. This hard clustering strategy allows us to train an LM that accounts for all the words rather than simply those NEs used in LDA analysis, as well as use higher order N-gram LMs, unlike the unigram based LDA in most previous work.

3.1.2. Clustering

We use the CLUTO toolkit [13] to perform clustering for the text collection. It finds a predefined number of clusters based on a specific criterion, for which we chose the following function (maximize the within-class similarity):

$$(S_1 S_2 ... S_K)^* = \operatorname{argmax} \sum_{i=1}^{K} \sqrt{\sum_{v,u \in S_i} sim(v,u)}$$

where K is the desired number of clusters, S_i is the set of documents belonging to the i^{th} cluster, v and u represent two documents, and $sim(v, u)$ is the similarity between them. We use the cosine distance to measure the similarity between two documents:

$$sim(u,v) = \frac{\overrightarrow{v} \times \overrightarrow{u}}{\|\overrightarrow{v}\| \times \|\overrightarrow{u}\|} \quad (1)$$

where $\overrightarrow{v}$ and $\overrightarrow{u}$ are the feature vectors representing the two documents respectively, again based on the NE hypotheses. The elements in every feature vector are also scaled based on their term frequency and inverse document frequency, a concept widely used in information retrieval. After clustering, we train an N-gram topic LM for each cluster using the documents in it.

3.2. Testing

During testing, a dynamically adaptive LM based on the topic analysis result is combined with the general LM to predict the probability of a word w_k given its history h_k, i.e.,

$$p(w_k|h_k) = \lambda * p_{general}(w_k|h_k) + (1-\lambda) * p_{adapt}(w_k|h_k) \quad (2)$$

where λ is the interpolation weight. The ways used to determine the new adapted LM (corresponding to p_{adapt} in the formula above) differ for the LDA and clustering based approaches.

3.2.1. LDA

For a test document $d = w_1, w_2, \ldots, w_n$ that is generated by multiple topics under the LDA assumption, we formulate a dynamically adapted topic model using the mixture of LMs from different topics:

$$p_{LDA-adapt}(w_k|h_k) = \sum_{i=1}^{K} \gamma_i \times p_{z_i}(w_k|h_k) \quad (3)$$

where $p_{z_i}(w_k|h_k)$ stands for the i^{th} topic LM, and γ_i is the mixture weight. Different from the idea of dynamic topic adaptation in [8], we propose a new weighting scheme to calculate γ_i that directly uses the two resulting matrices from LDA analysis during training:

$$\gamma_i = \sum_{j=1}^{n} p(z_i|w_j) p(w_j|d)$$

$$p(z_i|w_j) = \frac{f_{ji}}{\sum_{p=1}^{K} f_{jp}}$$

$$p(w_j|d) = \frac{freq(w_j|d)}{\sum_{q=1}^{n} freq(w_q|d)}$$

where $freq(w_j|d)$ is the frequency of a word w_j in the document d. Other notations are consistent with the previous definitions.

3.2.2. Clustering

In the clustering based approach, we use the cross entropy measure to determine the best topic for a test document, which was shown to outperform the cosine similarity distance measurement [5]. For a document $d = w_1, w_2, ..., w_n$, with a word distribution $p_d(w)$ and a cluster S with the associated topic specific LM $p_s(w)$, the cross entropy $CE(d, S)$ can be computed as the following using the unigram LM:

$$CE(d,S) = -\sum_{i=1}^{n} p_d(w_i) log_2(p_s(w_i)).$$

In other words, it is the perplexity of the test document d based on the LM corresponding to topic S. For the test document, we select the cluster S^* that yields the lowest perplexity:

$$S^* = \underset{1 \le i \le K}{\operatorname{argmin}}\, CE(d, S_i)$$

The LM corresponding to this topic S^* is then used in Eq (2) to combine with the generic LM.

4. N-BEST LIST RESCORING EXPERIMENTS

4.1. Data and Experimental Setup

The data set we used for N-best list rescoring is the GALE Mandarin 2007 Dev set. It contains about one hour of broadcast news (BN) (from 40 shows), and 1.5 hours of broadcast conversation (BC) speech (from 37 shows). The transcript has 2,000 utterance segments in this data set (46,819 characters).

We used a state-of-the-art Mandarin speech recognizer [14]. The recognizer's vocabulary consists of about 60K words. The baseline trigram LM was trained from a variety of data sources provided by LDC for the GALE project. There are about 700 millions words used for LM training. The recognizer uses a Gaussian Mixture model to identify the speech portion, and further segments them into short utterances. There are at most 1,000 hypotheses for each utterance segment. Our goal is to infer the "topic" information based on the hypotheses, and combine the acoustic scores and LM scores based on the new adapted LMs to rerank the hypotheses. The NE tagger we used is based on [15]. The followings are a few issues specific to our set up for the N-best list rescoring experiments.

- Training document generation
  Since the transcripts used to train the N-gram LMs do not have any topic annotation, for the purpose of topic analysis, we simply split them into shorter segments, each with 1,000 sentences, resulting in 40,378 segments in total. We used these as the "documents" for topic analysis (clustering or LDA) during training.
- NE pruning
  Automatic NE tagging is performed on the LM training data. Initially there were more than 4 million unique NE hypotheses on the entire training set. The NE tagger generates many false alarms, some of which are due to the style mismatch between the NE tagger training (well-written text) and testing (spoken language). In addition, this many of NE hypotheses make it computationally expensive for topic inference. To resolve these problems, we removed the NE hypotheses that have occurred fewer than a predefined threshold (100 in our experiments) in the training set. This yielded 29,310 unique NE hypotheses across all the documents. Pruning effectively reduces the "vocabulary" size for topic analysis, as well as removes many incorrect NE hypotheses.
- Testing document and segments
  For testing, we need to choose a segment to form a document for topic analysis for rescoring. We used the acoustic segments in the recognition output. It is generated mainly based on pause information, and thus may contain more than one sentence, or an incomplete sentence. Other possibilities include using automatic speaker segmentation or story/topic segmentation. Since those information is not readily available in the recognition output, and is also far from accurate, we leave the investigation of using different segments for the future work. For each acoustic segment, we use its recognition hypotheses (at most 1,000) to form a test document.

Table 1 summarizes the data set up for the rescoring experiments. The number of "documents" for training is from the sentence-based splitting, and for testing is the number of utterance segments from the recognizer. The number of NEs for testing is from the 1,000 recognition hypotheses, not from the reference transcription. There is no pruning of NE hypotheses during testing. However, the number of NEs shown in the table for testing corresponds to the NEs that have appeared in the training set. Only these will contribute to the computation of topic mixture weights in LDA-based adaptation. There are about 22K unique NE hypotheses in total from the N-best recognition hypotheses, but we do not need to consider all of them.

	training	testing
num. of "documents"	40,378	1,676
num. of words	700 million words	46,819 characters
num. of NEs after pruning	29,310	1,947

Table 1. Summary of data information in the large vocabulary Mandarin ASR task.

As discussed in Section 3.2, a mixture topic model is used in LDA-based LM adaptation, and a single topic is used in the clustering-based approach. For LDA-based adaptation, we used the NE hypotheses from the N-best list to find the LM mixture weights. For clustering-based adaptation using the cross-entropy criteria, we do not need to use the NE hypotheses during testing. Instead the entire recognition hypotheses are used to calculate cross entropy and find the best matched topic cluster. The adapted LMs are interpolated with the trigram LM, with an interpolation weight of 0.6 for these two approaches.

4.2. Rescoring Results

Recognition performance for Mandarin is measured using the character error rate (CER). Table 2 shows the N-best rescoring results (insertion, deletion, substitution errors, and the overall CER) on the GALE Mandarin dev set using the LDA and clustering-based LM adaptation approaches. For each setup, we present separate error rate for BN and BC respectively, and also the average error rate on the entire set. We evaluate two different numbers of topics (10 and 50 in this experiment) for the two approaches.

We can see that both clustering and LDA outperform the baseline trigram LM, yielding slightly lower error rate. The gain is observable for both BN and BC. In addition, clustering based topic modeling performs slightly better than LDA, unlike the perplexity results in [5]. For the two different number of topics, 10 and 50, we notice that a bigger number of topics degrades the rescoring performance in this experiment. This is also different from the perplexity results we obtained in [5], where increasing the number of topics helps. For LDA, this might be because of the number of mixture models we used in the current set up (9 used in the SRILM [16]). For the clustering-based approach with a single topic for LM adaptation, this might be explained by the smaller data size used to train the single topic adapted LM when the topic number increases. Further investigation about the topic numbers is needed.

Using a higher order LM (e.g., 5-gram LM) or class-based LMs on this data set yields a CER of 14.5%. Therefore using topic adaptation slightly outperforms those LMs. Given that the N-gram LM was trained from a large corpus, this improvement over the state-of-the-art recognition system is quite promising.

		Error rate (%)			
		sub	ins	del	CER
Baseline, 3-gram	BN	3.5	1.0	0.2	4.7
	BC	11.7	8.5	1.2	21.3
	Avg	8.3	5.4	0.8	**14.6**
LDA, 10 topics	BN	3.4	1.0	0.2	4.5
	BC	11.4	8.5	1.1	21.0
	Avg	8.1	5.5	0.7	**14.3**
LDA, 50 topics	BN	3.4	1.0	0.2	4.6
	BC	11.6	8.5	1.1	21.2
	Avg	8.3	5.5	0.7	**14.5**
Clustering, 10 topics	BN	3.3	1.0	0.2	4.5
	BC	11.2	8.4	1.2	20.8
	Avg	8.0	5.4	0.8	**14.2**
Clustering, 50 topics	BN	3.5	1.0	0.2	4.7
	BC	11.3	8.4	1.2	20.9
	Avg	8.1	5.4	0.8	**14.3**

Table 2. N-best list rescoring results (error rate %) using LDA and clustering-based LM adaptation on the GALE Mandarin dev set. 10 and 50 topics are used respectively for the two approaches.

One important advantage of using NE hypotheses is that it introduces new "words" that are not in the dictionary, allowing topic analysis to use more information than the existing words. In the recognizer used in the rescoring experiments, the vocabulary consists of about 60K words. Among the pruned NE hypotheses (29,310 NEs, see Table 1), only 12,311 NEs (42%) appear in the vocabulary, and the rest are new words. Therefore, this is an effective way to obtain multiwords.

5. CONCLUSIONS

We have investigated two topic modeling approaches for LM adaptation, LDA and clustering, both using NE information. For LDA, we also proposed a topic mixture model based on the word, document, and topic distribution. Unlike our previous work that only evaluated the new adaption approach using perplexity measurement, in this work we focus on N-best list rescoring for speech recognition. Our experiments have shown that using the topic adapted LM, the character error rate is improved slightly compared to the baseline trigram LM using a state-of-the-art recognition system. Though some errors of NE recognition may be introduced, our results indicate that exploring NEs for topic analysis is promising for LM adaptation. Between the two topic modeling approaches, we found the difference is rather small, with clustering achieving slightly better performance than LDA, an observation different from the perplexity results.

In our future work, we will identify appropriate segments, such as using speaker or story segmentation, to form test "documents" for N-best list rescoring. In addition, instead of using all the recognition utterance hypotheses and treating them equally to determine the topic information, we will investigate whether we can select more indicative NE hypotheses or use confidence measure associated with the ASR hypotheses. Finally we plan to perform a soft clustering in LDA-based training to allow each document to contribute to multiple topic LMs.

6. ACKNOWLEDGMENTS

We thank Mari Ostendorf, Mei-Yuh Hwang, Wen Wang, and Andreas Stolcke for their useful discussions and help with the rescoring experiments, and Heng Ji and Ralph Grishman for sharing the Mandarin named entity tagger. This work is supported by DARPA under Contract No. HR0011-06-C-0023. Distribution is unlimited. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

7. REFERENCES

[1] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models," in *Proc. of ICSLP*, 1996.

[2] J. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *IEEE Transactions on Speech and Audio Processing*, vol. 88(80), 2000.

[3] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proc. of Eurospeech*, 1999.

[4] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] F. Liu and Y. Liu, "Unsupervised language model adaptation incorporating named entity information," in *Proc. of ACL*, 2007.

[6] R. Kuhn and R.D. Mori, "A cache-based natural language model for speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 570–583, 1990.

[7] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer, Speech and Language*, vol. 10, pp. 187–228, 1996.

[8] Y.C. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference," in *Proc. of INTERSPEECH*, 2005.

[9] P. Hsu and J. Glass, "Style & topic language model adaptation using HMM-LDA," in *Proc. of EMNLP*, 2006, pp. 373–381.

[10] D. Mrva and P.C. Woodland, "Unsupervised language model adaptation for Mandarin broadcast conversation transcription," in *Proc. of INTERSPEECH*, 2006, pp. 2206–2209.

[11] A. Heidel, H. Chang, and L. Lee, "Language model adaptation using latent dirichlet allocation and an efficient topic inference algorithm," in *Proc. of Interspeech*, 2007.

[12] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum, "Integrating topics and syntax," *Adv. in Neural Information Processing Systems*, vol. 17, pp. 537–544, 2004.

[13] G. Karypis, "Software for clustering high-dimensional datasets," http://glaros.dtc.umn.edu/gkhome/views/cluto.

[14] M. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. of Interspeech*, 2006.

[15] H. Ji and R. Grishman, "Improving nametagging by reference resolution and relation detection," in *Proc. of ACL*, 2005, pp. 411–418.

[16] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. of ICSLP*, 2002.