Unsupervised Language Model Adaptation Incorporating Named Entity Information

Feifan Liu and Yang Liu
Department of Computer Science
The University of Texas at Dallas, Richardson, TX, USA
{ffliu,yangl}@hlt.utdallas.edu

Abstract

Language model (LM) adaptation is important for both speech and language processing. It 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, this paper investigates how effectively using named entity (NE) information, instead of considering all the words, helps LM adaptation. We evaluate two latent topic analysis approaches in this paper, namely, clustering and Latent Dirichlet Allocation (LDA). In addition, a new dynamically adapted weighting scheme for topic mixture models is proposed based on LDA topic analysis. Our experimental results show that the NE-driven LM adaptation framework outperforms the baseline generic LM. The best result is obtained using the LDA-based approach by expanding the named entities with syntactically filtered words, together with using a large number of topics, which yields a perplexity reduction of 14.23% compared to the baseline generic LM.

1 Introduction

Language model (LM) adaptation plays an important role in speech recognition and many natural language processing tasks, such as machine translation and information retrieval. Statistical N-gram LMs have been widely used; however, they capture only local contextual information. In addition, even with the increasing amount of LM training data, there is often a mismatch problem because of differences in domain, topics, or styles. Adaptation of LM, therefore, is very important in order to better deal with a variety of topics and styles.

Many studies have been conducted for LM adaptation. One method is supervised LM adaptation, where topic information is typically available and a topic specific LM is interpolated with the generic LM (Kneser and Steinbiss, 1993; Suzuki and Gao, 2005). In contrast, various unsupervised approaches perform latent topic analysis for LM adaptation. To identify implicit topics from the unlabeled corpus, one simple technique is to group the documents into topic clusters by assigning only one topic label to a document (Iyer and Ostendorf, 1996). 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) (Bellegarda, 2000), probabilistic latent semantic analysis (PLSA) (Gildea and Hofmann, 1999), and LDA (Blei et al., 2003). 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. Named entities are very common in domains such as newswire or broadcast news, and carry valuable information, which we hypothesize is topic indicative and useful for latent topic analysis. We compare different latent topic generation approaches as well as model adaptation methods, and propose an LDA based dynamic weighting method for the topic mixture model. Furthermore, we expand

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 672–679,
Prague, Czech Republic, June 2007.

named entities by incorporating other content words, in order to capture more topic information. Our experimental results show that the proposed method of incorporating named information in LM adaptation is effective. In addition, we find that for the LDA based adaptation scheme, adding more content words and increasing the number of topics can further improve the performance significantly.

The 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. Experimental results are presented and discussed in Section 4. Conclusion and future work appear in Section 5.

2 Related Work

There has been a lot of previous related work on LM adaptation. Suzuki and Gao (2005) compared different supervised LM adaptation approaches, and showed that three discriminative methods significantly outperform the maximum a posteriori (MAP) method. For unsupervised LM adaptation, an earlier attempt is a cache-based model (Kuhn and Mori, 1990), 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 (Rosenfeld, 1996).

Latent topic analysis has recently been investigated extensively for language modeling. Iyer and Ostendorf (1996) used hard clustering to obtain topic clusters for LM adaptation, where a single topic is assigned to each document. Bellegarda (2000) 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. For example, Gildea and Hofmann (1999) reported noticeable perplexity reduction via a dynamic combination of many unigram topic models with a generic trigram model. Proposed by Blei et al. (2003), 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 can overcome the drawbacks in the PLSA model, and has been shown to outperform PLSA in corpus perplexity and text classification experiments (Blei et al., 2003). Tam and Schultz (2005) 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 (2006) investigated using hidden Markov model with LDA to allow for both topic and style adaptation. Mrva and Woodland (2006) achieved 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.

In this paper, we investigate unsupervised LM adaptation using clustering and LDA based topic analysis. Unlike the clustering based interpolation method as in (Iyer and Ostendorf, 1996), we explore different distance measure methods for topic analysis. Different from the LDA based framework as in (Tam and Schultz, 2005), we propose a novel dynamic weighting scheme for the topic adapted LM. More importantly, the focus of our work is to investigate the role of named entity information in LM adaptation, which to our knowledge has not been explored.

3 Unsupervised LM Adaptation Integrating Named Entities (NEs)

3.1 Overview of the NE-driven LM Adaptation Framework

Figure 1 shows our unsupervised LM adaptation framework using NEs. For training, we use the text collection to train the generic word-based N-gram LM. Then we apply named entity recognition (NER) and topic analysis to train multiple topic specific N-gram LMs. During testing, NER is performed on each test document, and then a dynamically adaptive LM based on the topic analysis result is combined with the general LM. Note that in this figure, we evaluate the performance of LM adaptation using the perplexity measure. We will evaluate this framework for N-best or lattice rescoring in speech recognition in the future.

In our experiments, different topic analysis methods combined with different topic matching and adaptive schemes result in several LM adaptation paradigms, which are described below in details.

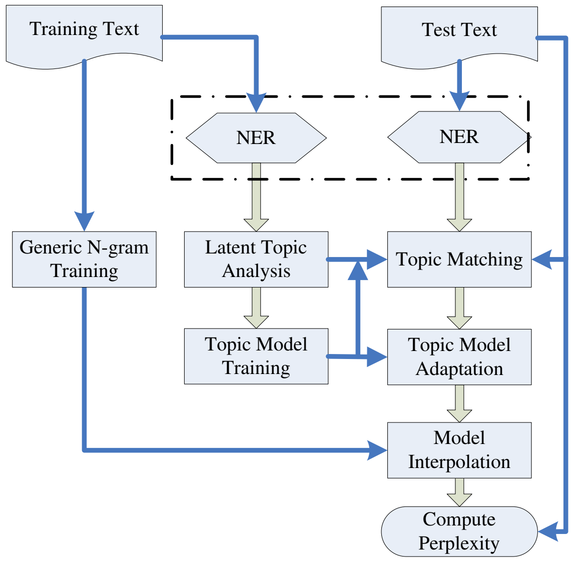

Figure 1. Framework of NE-driven LM adaptation.

3.2 NE-based Clustering for LM Adaptation

Clustering is a simple unsupervised topic analysis method. We use NEs to construct feature vectors for the documents, rather than considering all the words as in most previous work. We use the CLUTO[1] toolkit to perform clustering. It finds a predefined number of clusters based on a specific criterion, for which we chose the following function:

$$(S_1 S_2 \cdots S_k)^* = \arg\max \sum_{i=1}^{K} \sqrt{\sum_{v,u \in S_i} sim(v,u)}$$

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:

$$sim(v,u) = \frac{\vec{v} \cdot \vec{u}}{\|\vec{v}\| \cdot \|\vec{u}\|} \qquad (1)$$

where $\vec{v}$ and $\vec{u}$ are the feature vectors representing the two documents respectively, in our experiments composed of NEs. For clustering, the elements in every feature vector are scaled based on their term frequency and inverse document frequency, a concept widely used in information retrieval.

After clustering, we train an N-gram LM, called a topic LM, for each cluster using the documents in it.

During testing, we identify the 'topic' for the test document, and interpolate the topic specific LM with the background LM, that is, if the test document belongs to the cluster S^*, we can predict a word w_k in the document given the word's history h_k using the following equation:

$$\begin{aligned} p(w_k \mid h_k) &= \lambda p_{General}(w_k \mid h_k) \\ &\quad + (1-\lambda) p_{Topic-S^*}(w_k \mid h_k) \end{aligned} \qquad (2)$$

where λ is the interpolation weight.

We investigate two approaches to find the topic assignment S^* for a given test document.

(A) cross-entropy measure

For a test document $d=w_1,w_2,\ldots,w_n$ with a word distribution $p_d(w)$ and a cluster S with a topic LM $p_s(w)$, the cross entropy $CE(d, S)$ can be computed as:

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

From the information theoretic perspective, the cluster with the lower cross entropy value is expected to be more topically correlated to the test document. For each test document, we compute the cross entropy values according to different clusters, and select the cluster S^* that satisfies:

$$S^* = \arg\min_{1 \le i \le K} CE(d, S_i)$$

(B) cosine similarity

For each cluster, its centroid can be obtained by:

$$cv_i = \frac{1}{n_i} \sum_{k=1}^{n_i} u_{ik}$$

where u_{ik} is the vector for the k^{th} document in the i^{th} cluster, and n_i is the number of documents in the i^{th} cluster. The distance between the test document and a cluster can then be easily measured by the cosine similarity function as in Equation (1). Our goal here is to find the cluster S^* which the test document is closest to, that is,

$$S^* = \arg\max_{1 \le i \le K} \frac{\vec{d} \cdot \overrightarrow{cv_i}}{\|\vec{d}\| \cdot \|\overrightarrow{cv_i}\|}$$

[1] Available at http://glaros.dtc.umn.edu/gkhome/views/cluto

where $\vec{d}$ is the feature vector for the test document.

3.3 NE-based LDA for LM Adaptation

LDA model (Blei et al., 2003) has been introduced as a new, semantically consistent generative model, which overcomes overfitting and the problem of generating new documents in PLSA. It is a three-level hierarchical Bayesian model. Based on the LDA model, a document d is generated as follows.

- Sample a vector of K topic mixture weights θ from a prior Dirichlet distribution with parameter α:

$$f(\theta;\alpha) = \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- For each word w in d, pick a topic k from the multinomial distribution θ.
- Pick a word w from the multinomial distribution $\beta_{w,k}$ given the k^{th} topic.

For a document d=w_1,w_2,…w_n, the LDA model assigns it the following probability:

$$p(d) = \int_{\theta} \left(\prod_{i=1}^{n} \sum_{k=1}^{K} \beta_{w_i k} \cdot \theta_k \right) f(\theta;\alpha) d\theta$$

We use the MATLAB topic Toolbox 1.3 (Griffiths et al., 2004) in the training set to obtain the document-topic matrix, DP, and the word-topic matrix, WP. Note that here "words" correspond to the elements in the feature vector used to represent the document (e.g., NEs). 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,…,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,…,K) over the training set.

For training, we assign a topic z_i^* to a document d_i such that $z_i^* = \operatorname{argmax}_{1 \le k \le K} c_{ik}$. Based on the documents belonging to the different topics, 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 previous work.

For a test document d = w_1,w_2,…,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:

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

where $p_{z_i}(w_k \mid h_k)$ stands for the i^{th} topic LM, and γ_i is the mixture weight. Different from the idea of dynamic topic adaptation in (Tam and Schultz, 2005), we propose a new weighting scheme to calculate γ_i that directly uses the two resulting matrices from LDA analysis during training:

$$\gamma_k = \sum_{j=1}^{n} p(z_k \mid w_j) p(w_j \mid d)$$

$$p(z_k \mid w_j) = \frac{f_{jk}}{\sum_{p=1}^{K} f_{jp}}, \quad p(w_j \mid d) = \frac{freq(w_j)}{\sum_{q=1}^{n} freq(w_q)}$$

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

Then we interpolate this adapted topic model with the generic LM, similar to Equation (2):

$$\begin{aligned} p(w_k \mid h_k) = {} & \lambda p_{General}(w_k \mid h_k) \\ & + (1-\lambda) p_{LDA-adapt}(w_k \mid h_k) \end{aligned} \tag{3}$$

4 Experiments

4.1 Experimental Setup

	# of files	# of words	# of NEs
Training Data	23,985	7,345,644	590,656
Test Data	2,661	831,283	65,867

Table 1. Statistics of our experimental data.

The data set we used is the LDC Mandarin TDT4 corpus, consisting of 337 broadcast news shows with transcriptions. These files were split into small pieces, which we call documents here, according to the topic segmentation information marked in the LDC's transcription. In total, there are 26,646 such documents in our data set. We randomly chose 2661 files as the test data (which is balanced for different news sources). The rest was used for topic analysis and also generic LM training. Punctuation marks were used to determine sentences in the transcriptions. We used the NYU NE tagger (Ji and Grishman, 2005) to recognize four kinds of NEs: Person, Location, Organization, and Geo-political. Table 1 shows the statistics of the data set in our experiments.

We trained trigram LMs using the SRILM toolkit (Stolcke, 2002). A fixed weight (i.e., λ in Equation (2) and (3)) was used for the entire test set when interpolating the generic LM with the adapted topic LM. Perplexity was used to measure the performance of different adapted LMs in our experiments.

4.2 Latent Topic Analysis Results

	Topic	# of Files	Top 10 Descriptive Items (Translated from Chinese)
Clustering Based	1	3526	U.S., Israel, Washington, Palestine, Bush, Clinton, Gore, Voice of America, Mid-East, Republican Party
	2	3067	Taiwan, Taipei, Mainland, Taipei City, Chinese People's Broadcasting Station, Shuibian Chen, the Executive Yuan, the Legislative Yuan, Democratic Progressive Party, Nationalist Party
	3	4857	Singapore, Japan, Hong Kong, Indonesia, Asia, Tokyo, Malaysia, Thailand, World, China
	4	4495	World, German, Landon, Russia, France, England, Xinhua News Agency, Europe, U.S., Italy
	5	7586	China, Beijing, Nation, China Central Television Station, Xinhua News Agency, Shanghai, World, State Council, Zemin Jiang, Beijing City
LDA Based	1	5859	China, Japan, Hong Kong, Beijing, Shanghai, World, Zemin Jiang, Macao, China Central Television Station, Africa
	2	3794	U.S., Bush, World, Gore, South Korea, North Korea, Clinton, George Walker Bush, Asia, Thailand
	3	4640	Singapore, Indonesia, Team, Israel, Europe, Germany, England, France, Palestine, Wahid
	4	4623	Taiwan, Russia, Mainland, India, Taipei, Shuibian Chen, Philippine, Estrada, Communist Party of China, RUS.
	5	4729	Xinhua News Agency, Nation, Beijing, World, Canada, Sydney, Brazil, Beijing City, Education Ministry, Cuba

Table 2. Topic analysis results using clustering and LDA (the number of documents and the top 10 words (NEs) in each cluster).

For latent topic analysis, we investigated two approaches using named entities, i.e., clustering and LDA. 5 latent topics were used in both approaches. Table 2 illustrates the resulting topics using the top 10 words in each topic. We can see that the words in the same cluster share some similarity and that the words in different clusters seem to be 'topically' different. Note that errors from automatic NE recognition may impact the clustering results. For example, '队/team' in the table (in topic 3 in LDA results) is an error and is less discriminative for topic analysis.

Table 3 shows the perplexity of the test set using the background LM (baseline) and each of the topic LMs, from clustering and LDA respectively. We can see that for the entire test set, a topic LM generally performs much worse than the generic LM. This is expected, since the size of a topic cluster is much smaller than that of the entire training set, and the test set may contain documents from different topics. However, we found that when using an optimal topic model (i.e., the topic LM that yields the lowest perplexity among the 5 topic LMs), 23.45% of the documents in the test set have a lower perplexity value than that obtained from the generic LM. This suggests that a topic model could benefit LM adaptation and motivates a dynamic topic adaptation approach for different test documents.

	Perplexity
Baseline	502.02
CL-1	1054.36
CL-2	1399.16
CL-3	919.237
CL-4	962.996
CL-5	981.072
LDA-1	1224.54
LDA-2	1375.97
LDA-3	1330.44
LDA-4	1328.81
LDA-5	1287.05

Table 3. Perplexity results using the baseline LM vs. the single topic LMs.

4.3 Clustering vs. LDA Based LM Adaptation

In this section, we compare three LM adaptation paradigms. As we discussed in Section 3, two of them are clustering based topic analysis, but using different strategies to choose the optimal cluster; and the third one is based on LDA analysis that uses a dynamic weighting scheme for adapted topic mixture model.

Figure 2 shows the perplexity results using different interpolation parameters with the general LM. 5 topics were used in both clustering and LDA based approaches (as in Section 4.2). "CL-CE" means clustering based topic analysis via cross entropy criterion, "CL-Cos" represents clustering based topic analysis via cosine distance criterion, and "LDA-MIX" denotes LDA based topic mixture model, which uses 5 mixture topic LMs.

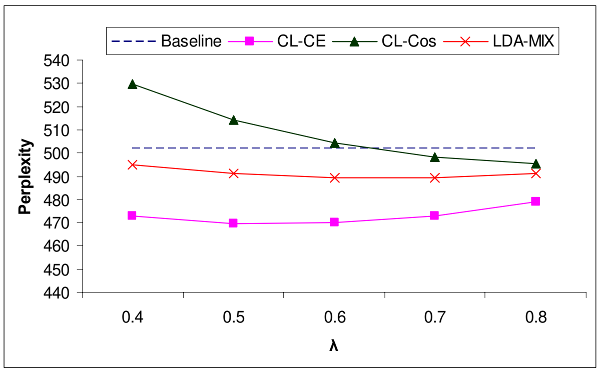

Figure 2. Perplexity using different LM adaptation approaches and different interpolation weights λ with the general LM.

We observe that all three adaptation approaches outperform the baseline when using a proper interpolation weight. "CL-CE" yields the best perplexity of 469.75 when λ is 0.5, a reduction of 6.46% against the baseline perplexity of 502.02. For clustering based adaptation, between the two strategies used to determine the topic for a test document, "CL-CE" outperforms "CL-Cos". This indicates that the cosine distance measure using only names is less effective than cross entropy for LM adaptation. In addition, cosine similarity does not match perplexity as well as the CE-based distance measure. Similarly, for the LDA based approach, using only NEs may not be sufficient to find appropriate weights for the topic model. This also explains the bigger interpolation weight for the general LM in CL-Cos and LDA-MIX than that in "CL-CE".

For a fair comparison between the clustering and LDA based LM adaptation approaches, we also evaluated using the topic mixture model for the clustering based approach and using only one topic in the LDA based method. For clustering based adaptation, we constructed topic mixture models using the weights obtained from a linear normalization of the two distance measures presented in Section 3.2. In order to use only one topic model in LDA based adaptation, we chose the topic cluster that has the largest weight in the adapted topic mixture model (as in Sec 3.3). Table 4 shows the perplexity for the three approaches (CL-Cos, CL-CE, and LDA) using the mixture topic models versus a single topic LM. We observe similar trends as in Figure 2 when changing the interpolation weight λ with the generic LM; therefore, in Table 4 we only present results for one optimal interpolation weight.

	Single-Topic	Mixture-Topic
CL-Cos (λ =0.7)	498.01	497.86
CL-CE (λ =0.5)	469.75	483.09
LDA (λ =0.7)	488.96	489.14

Table 4. Perplexity results using the adapted topic model (single vs. mixture) for clustering and LDA based approaches.

We can see from Table 4 that using the mixture model in clustering based adaptation does not improve performance. This may be attributed to how the interpolation weights are calculated. For example, only names are used in cosine distance, and the normalized distance may not be appropriate weights. We also notice negligible difference when only using one topic in the LDA based framework. This might be because of the small number of topics currently used. Intuitively, using a mixture model should yield better performance, since LDA itself is based on the assumption of generating words from multiple topics. We will investigate the impact of the number of topics on LM adaptation in Section 4.5.

4.4 Effect of Different Feature Configurations on LM Adaptation

We suspect that using only named entities may not provide enough information about the 'topics' of the documents, therefore we investigate expanding the feature vectors with other words. Since generally content words are more indicative of the topic of a document than function words, we used a POS tagger (Hillard et al., 2006) to select words for latent topic analysis. We kept words with three POS classes: noun (NN, NR, NT), verb (VV), and modifier (JJ), selected from the LDC POS set[2]. This is similar to the removal of stop words widely used in information retrieval.

Figure 3 shows the perplexity results for three different feature configurations, namely, all-words (w), names (n), and names plus syntactically filtered items (n+), for the CL-CE and LDA based approaches. The LDA based LM adaptation paradigm supports our hypothesis. Using named information instead of all the words seems to efficiently eliminate redundant information and achieve better performance. In addition, expanding named entities with syntactically filtered items yields further improvement. For CL-CE, using named information achieves the best result among the three configurations. This might be because that the clustering method is less powerful in analyzing the principal components as well as dealing with redundant information than the LDA model.

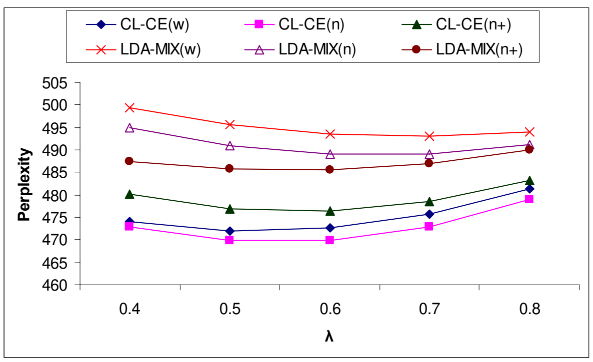

Figure 3. Comparison of perplexity using different feature configurations.

4.5 Impact of Predefined Topic Number on LM Adaptation

LDA based topic analysis typically uses a large number of topics to capture the fine grained topic space. In this section, we evaluate the effect of the number of topics on LM adaptation. For comparison, we evaluate this for both LDA and CL-CE, similar to Section 4.3. We use the "n+" feature configuration as in Section 4.4, that is, names plus POS filtered items. When using a single-topic adapted model in the LDA or CL-CE based approach, finer-grained topic analysis (i.e., increasing the number of topics) leads to worse performance mainly because of the smaller clusters for each topic; therefore, we only show results here using the mixture topic adapted models. Figure 4 shows the perplexity results using different numbers of topics. The interpolation weight λ with the general LM is 0.5 in all the experiments. For the topic mixture LMs, we used a maximum of 9 mixtures (a limitation in the current SRILM toolkit) when the number of topics is greater than 9.

We observe that as the number of topics increases, the perplexity reduces significantly for LDA. When the number of topics is 50, the adapted LM using LDA achieves a perplexity reduction of 11.35% compared to using 5 topics, and 14.23% against the baseline generic LM. Therefore, using finer-grained multiple topics in dynamic adaptation improves system performance. When the number of topics increases further, e.g., to 100, the performance degrades slightly. This might be due to the limitation of the number of the topic mixtures used. A similar trend is observable for the CL-CE approach, but the effect of the topic number is much greater in LDA than CL-CE.

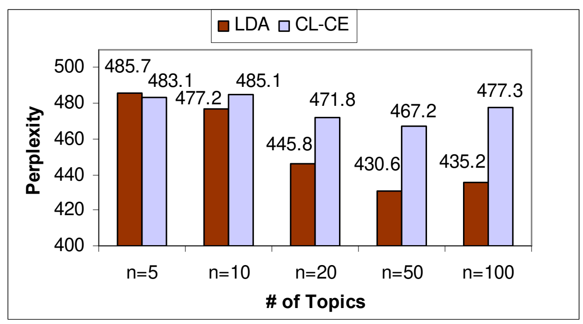

Figure 4. Perplexity results using different predefined numbers of topics for LDA and CL-CE.

4.6 Discussion

As we know, although there is an increasing amount of training data available for LM training, it is still only for limited domains and styles. Creating new training data for different domains is time consuming and labor intensive, therefore it is very important to develop algorithms for LM adaptation. We investigate leveraging named entities in the LM adaptation task. Though some errors of NER may be introduced, our experimental results have shown that exploring named information for topic analysis is promising for LM adaptation.

Furthermore, this framework may have other advantages. For speech recognition, using NEs for topic analysis can be less vulnerable to recognition

[2] See http://www.cis.upenn.edu/~chinese/posguide.3rd.ch.pdf

errors. For instance, we may add a simple module to compute the similarity between two NEs based on the word tokens or phonetics, and thus compensate the recognition errors inside NEs. Whereas, word-based models, such as the traditional cache LMs, may be more sensitive to recognition errors that are likely to have a negative impact on the prediction of the current word. From this point of view, our framework can potentially be more robust in the speech processing task. In addition, the number of NEs in a document is much smaller than that of the words, as shown in Table 1; hence, using NEs can also reduce the computational complexity, in particular in topic analysis for training.

5 Conclusion and Future Work

We compared several unsupervised LM adaptation methods leveraging named entities, and proposed a new dynamic weighting scheme for topic mixture model based on LDA topic analysis. Experimental results have shown that the NE-driven LM adaptation approach outperforms using all the words, and yields perplexity reduction compared to the baseline generic LM. In addition, we find that for the LDA based method, adding other content words, combined with an increased number of topics, can further improve the performance, achieving up to 14.23% perplexity reduction compared to the baseline LM.

The experiments in this paper combine models primarily through simple linear interpolation. Thus one direction of our future work is to develop algorithms to automatically learn appropriate interpolation weights. In addition, our work in this paper has only showed promising results in perplexity reduction. We will investigate using this framework of LM adaptation for N-best or lattice rescoring in speech recognition.

Acknowledgements

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

References

J. Bellegarda. 2000. Exploiting Latent Semantic Information in Statistical Language Modeling. In IEEE Transactions on Speech and Audio Processing. 88(80):1279-1296.

D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. 3:993-1022.

D. Gildea and T. Hofmann. 1999. Topic-Based Language Models using EM. In Proc. of Eurospeech.

T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. 2004. Integrating Topics and Syntax. Adv. in Neural Information Processing Systems. 17:537-544.

D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-Tur, M. Harper, M. Ostendorf, and W. Wang. 2006. Impact of Automatic Comma Prediction on POS/Name Tagging of Speech. In Proc. of the First Workshop on Spoken Language Technology (SLT).

P. Hsu and J. Glass. 2006. Style & Topic Language Model Adaptation using HMM-LDA. In Proc. of EMNLP, pp:373-381.

R. Iyer and M. Ostendorf. 1996. Modeling Long Distance Dependence in Language: Topic Mixtures vs. Dynamic Cache Models. In Proc. of ICSLP.

H. Ji and R. Grishman. 2005. Improving NameTagging by Reference Resolution and Relation Detection. In Proc. of ACL. pp: 411-418.

R. Kneser and V. Steinbiss. 1993. On the Dynamic Adaptation of Stochastic language models. In Proc. of ICASSP, Vol 2, pp: 586-589.

R. Kuhn and R.D. Mori. 1990. A Cache-Based Natural Language Model for Speech Recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 12: 570-583.

D. Mrva and P.C. Woodland. 2006. Unsupervised Language Model Adaptation for Mandarin Broadcast Conversation Transcription. In Proc. of INTERSPEECH, pp:2206-2209.

R. Rosenfeld. 1996. A Maximum Entropy Approach to Adaptive Statistical Language Modeling. Computer, Speech and Language, 10:187-228.

A. Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In Proc. of ICSLP.

H. Suzuki and J. Gao. 2005. A Comparative Study on Language Model Adaptation Techniques Using New Evaluation Metrics, In Proc. of HLT/EMNLP.

Y.C. Tam and T. Schultz. 2005. Dynamic Language Model Adaptation Using Variational Bayes Inference. In Proc. of INTERSPEECH, pp:5-8.