Novel Weighting Scheme for Unsupervised Language Model Adaptation Using Latent Dirichlet Allocation

Md. Akmal Haidar and Douglas O'Shaughnessy

Interspeech 2010

Pei-ning Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

Outline

- Introduction
- Latent dirichlet allocation
- Topic clustering and language model generation
- Language model adaptation
- Experiments and results
- Conclusions

Introduction

- Language model (LM) adaptation is required when the styles, domains or topics of the test data are mismatched with the training data.
- An adaptive language model seeks to maintain an adequate representation of the domain under changing conditions involving potential variations in vocabulary, content, syntax and style

Introduction (cont.)

- The idea of an unsupervised LM adaptation approach is to extract the latent topics from the training set and then adapt the topic specific LM with proper mixture weights, finally interpolated with the generic n-gram LM.
- There are various techniques to extract the latent semantic information from a training corpus such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) , and LDA

Introduction (cont.)

- LDA can be used to model an unseen document as it imposes a Dirichlet distribution over topic mixture weights corresponding to the documents in the corpus. However, the LDA model can be viewed as a mixture of unigram latent topic models.
- They propose the idea that the weights of topic models are generated using the word count of the topics generated by a hard-clustering method instead of using the LDA latent topic word count

Latent dirichlet allocation

- Each document $d = w_1, \ldots, w_n$ is generated as a mixture of unigram models, where the topic mixture weight θ is drawn from a prior Dirichlet distribution:
$$f\left(\theta;\alpha\right) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$
- For each word in document *d:*
  - Choose a topic k from the multinomial distribution $\theta(d)$
  - Choose a word w from the multinomial distribution $\Phi(w \mid k, \beta)$

Latent dirichlet allocation(cont.)

- where $\alpha = \{\alpha_1, \cdots, \alpha_k\}$ is used as the representation count for the K latent topics,
- θ indicates the relative importance of topics for a document
- $\Phi(w \mid k, \beta)$ represents the word probabilities conditioned on the topic with a Dirichlet prior and indicates the relative importance of particular words in a topic

- As a bag-of-word generative model, LDA assigns the following probability to a document $d = w_1, \ldots, w_n$ as:

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

Topic clustering and language model generation

- They have used the MATLAB topic modeling toolbox to get:
  - the word-topic matrix, WP,
  - the document-topic matrix,DP, using LDA.
    - WP(j,k) represents the number of times word w_j has been assigned to topic z_k over the training set.
    - DP(i,k) contains the counts of words in document d_i that are from a topic z_k (k=1,2…,K).
- For training, topic clusters are formed by assigning a topic z_i^* to a document d_i as:

$$z_i^* = \arg\max_{1 \le k \le K} DP(i,k)$$

Language model adaptation

- A document can be generated by a mixture of topics. So, for a test document $d = w_1, \ldots, w_n$, we can create a dynamically adapted topic model by using a mixture of LMs from different topics as: $P_{LDA-adapt}(w_k \mid h_k) = \sum_{i=1}^{K} \gamma_i p_{z_i}(w_k \mid h_k)$

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

$$P(z_k \mid w_j) = \frac{TF(j,k)}{\sum_{p=1}^{K} TF(j,p)}$$

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

$$P(w_k \mid h_k) = \lambda * P_{general}(w_k \mid h_k) + (1-\lambda) * P_{LDA-adapt}(w_k \mid h_k)$$

Experiments and results

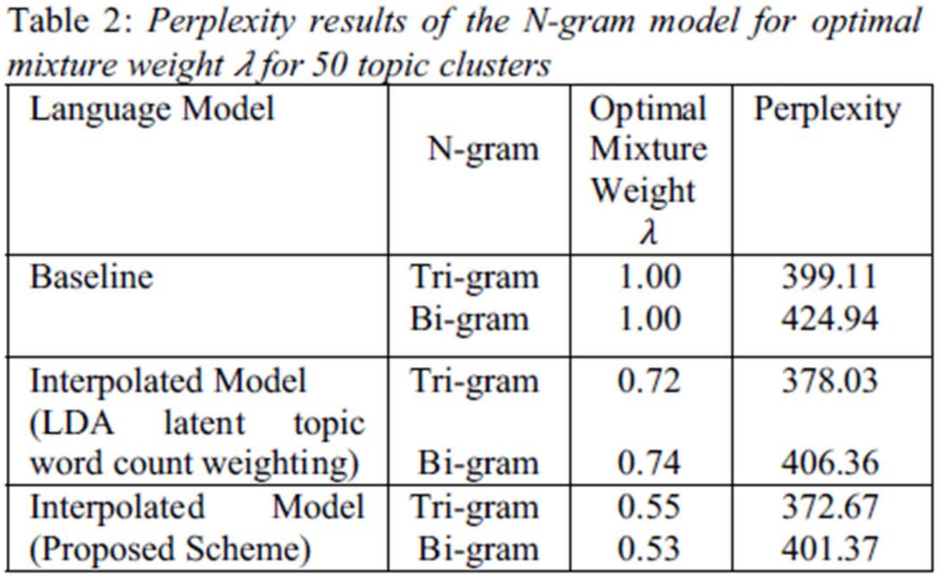

Table 2: *Perplexity results of the N-gram model for optimal mixture weight λ for 50 topic clusters*

Language Model	N-gram	Optimal Mixture Weight λ	Perplexity
Baseline	Tri-gram	1.00	399.11
	Bi-gram	1.00	424.94
Interpolated Model (LDA latent topic word count weighting)	Tri-gram	0.72	378.03
	Bi-gram	0.74	406.36
Interpolated Model (Proposed Scheme)	Tri-gram	0.55	372.67
	Bi-gram	0.53	401.37

- – They used perplexity and WER to measure the performance of their experiments here.

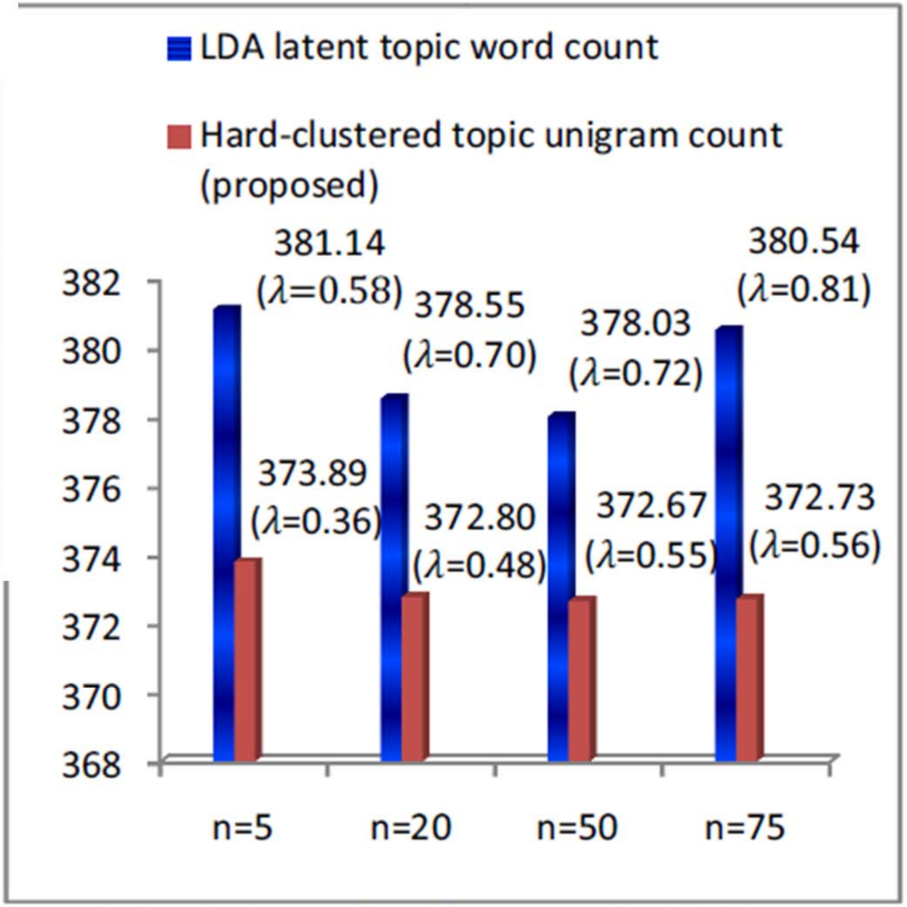

Figure 1: *Perplexity results of the test set for LDA latent topic word count and the proposed scheme using different numbers of topics and optimal mixture weights of the tri-gram model.*

Experiments and results(cont.)

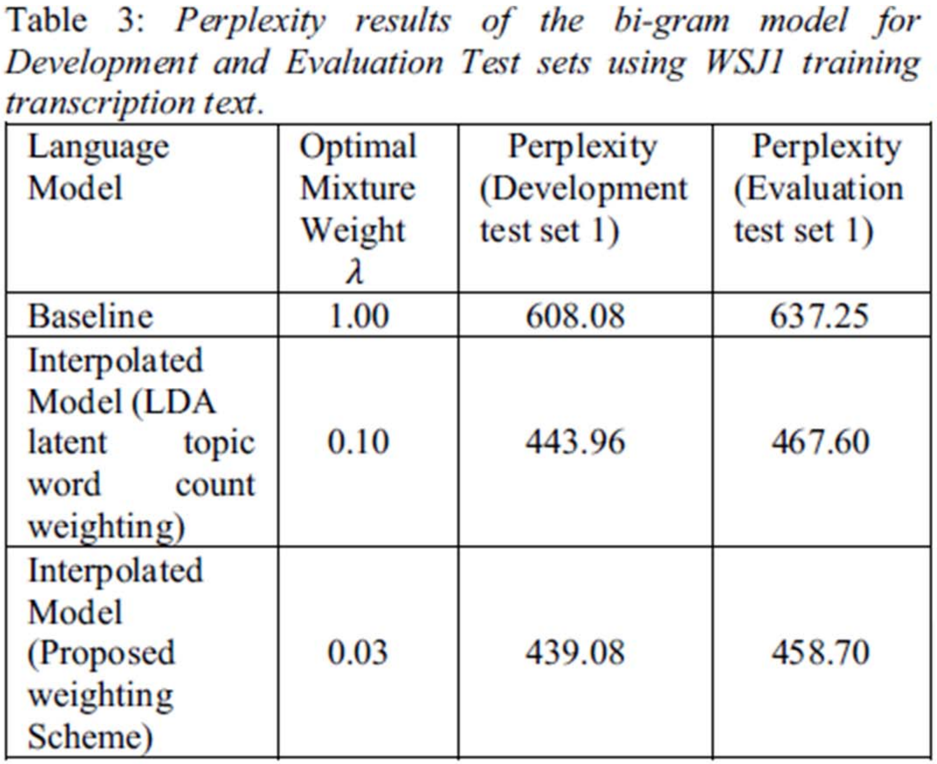

Table 3: *Perplexity results of the bi-gram model for Development and Evaluation Test sets using WSJ1 training transcription text.*

Language Model	Optimal Mixture Weight λ	Perplexity (Development test set 1)	Perplexity (Evaluation test set 1)
Baseline	1.00	608.08	637.25
Interpolated Model (LDA latent topic word count weighting)	0.10	443.96	467.60
Interpolated Model (Proposed weighting Scheme)	0.03	439.08	458.70

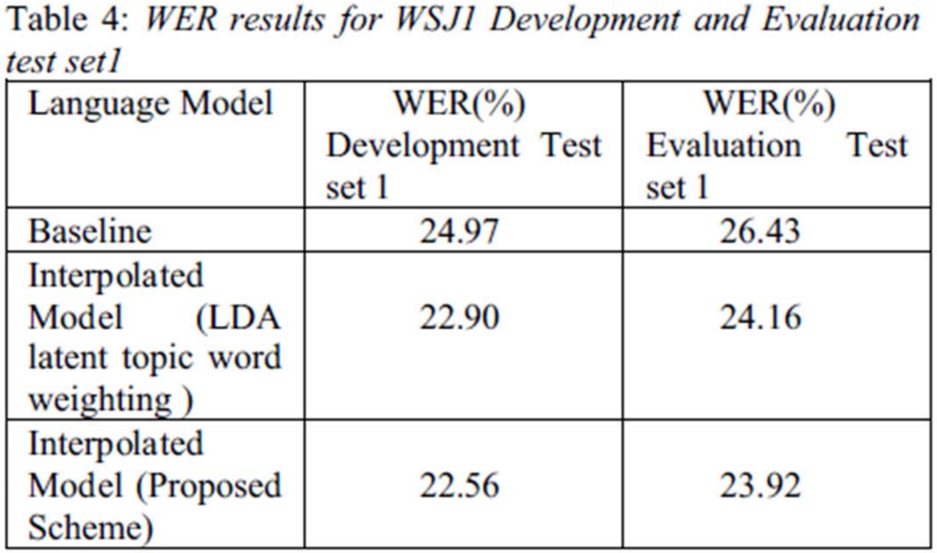

Table 4: *WER results for WSJ1 Development and Evaluation test set1*

Language Model	WER(%) Development Test set 1	WER(%) Evaluation Test set 1
Baseline	24.97	26.43
Interpolated Model (LDA latent topic word weighting)	22.90	24.16
Interpolated Model (Proposed Scheme)	22.56	23.92

Conclusions

- They proposed a novel weighting scheme for unsupervised language model adaptation using LDA and compared two weighting schemes for topic model adaption.
- The proposed word count weighting of the topic generated by hard clustering outperforms the LDA latent topic word count weighting in both perplexity and WER measurement.