Search and Classification Based Language Model Adaptation

Qin Shi$^1$, Stephen M. Chu$^2$, Wen Liu$^1$, Hong-Kwang Kuo$^2$, Yi Liu$^1$, Yong Qin$^1$

$^1$ IBM China Research Lab
$^2$ IBM T. J. Watson Research Center

{shiqin, liuwen, liuyyi, qinyong}@cn.ibm.com, {schu, hkuo}@us.ibm.com

Abstract

Adaptation techniques in language modeling have shown growing potentials in improving speech recognition performance. For topic adaptation, a set of pre-defined topic-specific language models are typically used, and adaptation is achieved through adjusting the interpolation weights. However, mismatch between the test data and the pre-defined models inevitably exists and is left untreated in the static approach. Instead of tuning the parameters in the existing models, this paper describes a method that dynamically extracts relevant documents from training sources according to intermediate decoding hypotheses to build new targeted language models. Different from general search-based document collection, a new and effective ranking method is used here for candidate extraction. The targeted language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-art Mandarin broadcast transcription system, and evaluated on the GALE task. We show that static topic language models are interpolated with the static topic language models and a general language model, and used for lattice rescoring. The proposed adaptation technique is implemented in a state-of-the-
combining dynamic LMs with static LM, interpolation techniques are widely used.

In this work, we focus on topic adaptation and introduce the search-based dynamic LM in the interpolation mix. In particular, the knowledge of the characteristics of the ASR output is exploited to improve the document ranking method in the search process.

2.1 SVM based corpus classification

Topic based language modeling is based on the assumption that the usage of content words typically depends on the topic of the text. A key step in constructing topic-specific LMs is to cluster the corpus into the appropriate topics.

The SVM, which has been shown to yield good generalization performance on a wide variety of classification problems with large-scale input space, e.g. handwritten character recognition, face detection, and text categorization, is employed in our system to carry out text classification and topic clustering.

Manually labeled documents are used to train the SVM. The feature vector representing each training sample is a vector of terms given by our Mandarin text segmenter. To reduce the number of nuisance features, words occurring in fewer than three documents are omitted, and stop words are deleted from the feature vector. The overall classification accuracy of the topic classifier as measured by the $F_1$ measure is 0.8. Based on the SVM classifier, the entire LM training data was partitioned into 129 topics, and 129 static topic LMs are built based on the clustered training documents.

2.2 Search based LM

Although static topic language models can help to improve recognition accuracy through modeling different distributions of content words among topics, such topic models might still be too general for a specific story. For instance, under the broad topic class of politics, there are many sub-topics that are not related to the current story. In order to make sure that the $n$-grams related to the story are promoted, the ideal scenario is that we can find documents whose content is closely matched to the specific story.

The ranking algorithm used in the search engine is a combination of vector space model (VSM), a common information retrieval method, and the Boolean model to determine how relevant a given document is to a user's query. The motivation of VSM is that the more times a query term appears in a document relative to the number of times the term appears in all the documents in the collection, the more relevant that document is to the query. The Boolean model is used to first narrow down the documents that need to be scored based on the use of Boolean logic in the query specification.

Note that our purpose of improving ASR performance is different from the general search requirements for information retrieval. The first difference is that there are errors in the ASR transcription. Such errors will misguide the search engine. The second difference is that the search criteria should be the similarity at the document level. Based on the above analysis, the keys for searching documents are carefully selected, and the scoring method is modified accordingly.

A topic $T_j$ is composed of utterances: $\{u_1, u_2, ..., u_n\}$, where an utterance $u_i$ is expressed as: $u_i = (w_{i1}, w_{i2}, ..., w_{im})$, and $w_j$ represents a word. Silence and hesitation words in the utterance are tagged as potential chop points, thus $u_i$ can be decomposed into word segments at these boundaries. After discarding very short segments, the word list is converted into a search key list $(key_{1j}, key_{2j}, ..., key_{mj})$. For each search key $key_{kj}$, a large number of documents are typically retrieved, which are first sorted by the ranking score, and $M$ documents with the highest scores are kept. After collecting all the documents selected by the keys in one topic, we define another score called the hit-rate score for each document, which is defined as the number of keys in the topic that occur in the document. According to this score, the documents are sorted again. Intuitively, the more similar the document is to the topic, the higher the document’s hit-rate score would be. The hit-rate score also helps to reduce/prevent irrelevant documents from being extracted due to ASR errors in the search keys. Finally, a given number of documents with the highest hit-rate scores are collected to build a 4-gram LM. We shall refer this LM as the search-based topic LM.

![Fig. 1](image1)

Fig. 1 The documents selection algorithm uses the ranking score and the hit-rate score to select and prune the relevant documents for building the search-based topic LM.

2.3 Interpolation method

When decoding a segment of speech that is assumed to be of a single topic, three kinds of language models can be combined:

![Fig. 2](image2)

Fig. 2 The proposed LM adaptation algorithm interpolates static-topic LMs, the search-based topic LM, and the general baseline LM for lattice rescoring.
interpolate all the models, not all static topic language models are involved. Using the perplexity of the first pass decoding result with respect to different topic LMs, a subset of the static topic LMs with the lowest perplexity scores are selected as the candidates to interpolate with other models. The interpolation weights are chosen to optimize perplexity based on the first pass decoding result.

Fig. 2 shows the entire search and classification based topic adaptation process.

3. System Description

3.1 System Architecture

The Mandarin broadcast transcription system [9] developed at IBM for the DARPA GALE project is shown in Fig 3. There are five main stages in the decoding pipeline: audio segmentation, speaker clustering, speaker independent (SI) decoding, and lattice rescoring.

In the Front-End, PLP features are used for segmentation and recognition. LDA and MLLT (maximum likelihood linear transform) are applied. An HMM-based classifier is used in the segmentation step. After segmentation and discarding the non-speech parts, the speech frames are clustered iteratively according to the feature distance among the segments.

SI decoding is performed first. Here three-state, left-to-right HMMs are used to represent 162 phones. The HMM states are context-dependent and clustered into equivalence classes by using decision trees. The distributions of 15K states are modeled by a pool of 500K Gaussian densities. The SI decoding output is used for speaker adaptation and SA decoding is subsequently carried out. In this step, a model-space method, MLLR and two feature-space methods, VTLN and fMLLR (feature/constrained MLLR) are used. In addition to the speaker adaptation procedures, discriminative training with minimum phone error (MPE) [10] and feature-space MPE (fMPE) are employed in the SA stage.

A general 4-gram LM is used in the SI and SA decoding stages. The model is generated by interpolated back-off 4-gram models and is smoothed by modified Kneser-Ney smoothing. The interpolation weights are chosen to optimize the perplexity of a held-out data set. To fully utilize the hypotheses generated in the decoding stage and the text database, the proposed LM adaptation technique is applied in the final lattice rescoring step.

3.2 Training Data

The acoustic training data is obtained from the LDC. The system is trained on a total of 1400 hours of audio. Manual transcriptions exist for 700 hours; the rest of the data was used through lightly supervised training.

The vocabulary has 107.8K words and for language model training data, about 1.76 G words from LDC and internet web data are used, including GIGA2, GaleY1Q1Q2, GaleY1Q3Q4, TDT2-5, TREC, and shared Web news text data. All data is dated before October 31, 2006. The LM training data was separated into 20 classes according to the data source, time, and style. A total 20 language models are separately built. These LMs are interpolated according to the held out set to give the baseline LM with 5.9M n-grams.

3.3 Lattice Rescoring and Topic Adaptation

In addition to the general LM used in SI and SA decoding, a new topic-adapted LM for lattice rescoring is obtained by interpolating the static topic LMs, the search-based topic LM, and the general LM.

For the static topic LMs, the topics are organized as a manually constructed tree with 129 leaf nodes. The topics include: politics, finance, sports, science, society, and so on. More than 20,000 Chinese news articles related to the topics are collected and annotated. Based on the SVM classifier, the entire LM training data was partitioned into 129 topics. And 129 language models are built based on the documents belong to the topics.

For the search-based topic LM, an off-line database with data collected before the GALE specified cut-off date (Oct. 31, 2006) is constructed for search in decoding time.

4. Experiments

4.1 Experimental Setup

Two GALE data sets are used to evaluate the Mandarin broadcast transcription system. The first is a community-defined development set, dev07, which contains 153 minutes and is chopped into 124 snippets (here each snippet represents a topic). The second is the 2007 evaluation set, eval07, which 142 minutes audio and is chopped into 119 snippets.

After SA decoding, LM rescoring is applied to the resulting lattices. The following five methods are compared:

bigLM:
LM rescoring is based on one big language model, composed of an interpolation of 20 sub-LMs. The entire set of snippets share the same language model, and the 1-best decoding result of the whole test set is used to determine the optimal interpolation weights.

topicLM:
Different from the bigLM, the snippets will use different interpolated LMs. For each snippet, an interpolated LM of four parts is constructed as follows. Three static topic LMs with the lowest perplexities for the first-pass decoded text are se-
lected. The three LMs are interpolated with the baseline general LM, with the interpolation weights optimized with respect to the first-pass decoded text.

**topicLM**: Different from **topicLM**, considering the data size limitation for broadcast conversation, all of the broadcast conversation data is used to build an additional topic LM. A total of 130 static topic LMs are involved in the candidate selection, among which four with the lowest perplexities for the first-pass decoded text are used in rescoring.

**localSearch_topicLM**:
LM rescoring is based on interpolating the baseline language model, four static topic LMs, and one search-based topic LM. To generate the search-based LM, we kept about 800 documents with the highest search ranking score for each search key in the snippet. All the documents retrieved by all key words in the snippet are collected; and only 20 documents with the highest hit-rate scores are kept to build the search-based topic LM.

**webSearch_topicLM**:
LM rescoring used the same way as in **localSearch_topicLM**. The difference is that the source database is expanded to include the web.

### 4.2 Results
The first-pass decoding result and the results of the four LM rescoring methods are shown in Table 1. The results are given in character error rates (CER).

<table>
<thead>
<tr>
<th>Method</th>
<th>dev07</th>
<th>eval07</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>11.5</td>
<td>10.2</td>
</tr>
<tr>
<td>bigLM</td>
<td>11.2</td>
<td>9.9</td>
</tr>
<tr>
<td>topicLM</td>
<td>11.2</td>
<td>9.7</td>
</tr>
<tr>
<td>topicLM+</td>
<td>10.9</td>
<td>9.5</td>
</tr>
<tr>
<td>localSearch_topicLM+</td>
<td>10.7</td>
<td>9.2</td>
</tr>
<tr>
<td>webSearch_topicLM+</td>
<td>10.3</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Compare to the baseline, the **localSearch_topicLM+** method achieves an impressive 1.0 absolute improvement in CER on eval07. Furthermore, the web search-based approach gives an additional 0.5 to attain an overall 1.5 point reduction, which represents a 14.7% relative improvement.

LM rescoring without topic adaptation (**bigLM**) accounts for 2.9% in the 14.7%, while the rest of the gain can be attributed to the combination of LM adaptation techniques described in this paper.

Note that the web search method was not used in the actual GALE evaluation due to rules of the program. Nonetheless, the experiment confirms that better hypotheses in fact exist in the input lattices, and in a practical broadcast audio transcription system, incorporating web search in LM adaptation may well be a very promising way to enhance recognition accuracy.

### 5. Conclusions
In this paper, we considered topic adaptation in language modeling. An effective search and classification based LM adaptation pipeline was proposed and implemented in a state-of-the-art ASR system. Evaluation on broadcast transcription tasks showed that the described technique lead to significant reductions in recognition errors. In the scope of this work, we have kept the focus on content similarity. The style of speech actually gives another interesting avenue to apply LM adaptation, especially for broadcast conversation contents. This direction will be explored in our future work.

### 6. References