NONLINEAR INTERPOLATION OF TOPIC MODELS FOR LANGUAGE MODEL ADAPTATION

Kristie Seymore, Stanley Chen and Ronald Rosenfeld

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

Topic adaptation for language modeling is concerned with adjusting the probabilities in a language model to better reflect the expected frequencies of topical words for a new document. The language model to be adapted is usually built from large amounts of training text and is considered representative of the current domain. In order to adapt this model for a new document, the topic (or topics) of the new document are identified. Then, the probabilities of words that are more likely to occur in the identified topic(s) than in general are boosted, and the probabilities of words that are unlikely for the identified topic(s) are suppressed.

We present a novel technique for adapting a language model to the topic of a document, using a nonlinear interpolation of n-gram language models. A three-way, mutually exclusive division of the vocabulary into *general*, *on-topic* and *off-topic* word classes is used to combine word predictions from a topic-specific and a general language model. We achieve a slight decrease in perplexity and speech recognition word error rate on a Broadcast News test set using these techniques. Our results are compared to results obtained through linear interpolation of topic models.

1. INTRODUCTION

A language model furnishes the probability $p(w|h)$ of a word w occurring given the previously occurring words, or history h. Language model adaptation deals with changing the probabilities of certain words from some set of initial values due to additional knowledge about the text under consideration. In topic adaptation, the topic(s) of a sample of text are identified and that information is used to adjust the probabilities of topical words in the model.

Topical words are those words whose frequencies depend strongly on topic. A topic-adapted language model should ideally assign a higher overall likelihood to new text than the initial model by increasing the probabilities of words it expects to encounter in the identified topic (*on-topic* words), and decreasing the probabilities of words that do not normally occur in the identified topic (*off-topic* words). The probabilities of non-topical, or *general*, words may not change at all, because they are equally likely for any topic. This paper introduces the notion of nonlinearly interpolating the predictions from a general and a topic-specific language model to boost the probabilities of on-topic words and suppress the probabilities of off-topic words.

Previous work in topic adaptation [1, 3, 4, 5, 7, 10, 11] has mainly focused on identifying topic-specific subsets of the training text and building language models from them. The topic language models are linearly interpolated with a general language model built from all of the training text. Using this technique, all of the available models are consulted for each word prediction, and interpolation weights λ_i govern how strongly each models' predictions are counted in the overall probability calculation, *i.e.*,

$$p_{\text{adapted}}(w \mid h) = \sum_i \lambda_i p_i(w \mid h) \qquad (1)$$

where the p_i denote the models being combined.

Nonlinear interpolation chooses, for each word in the vocabulary, the one model that is "most qualified" to provide the probabilities for that word. A model trained on all available data has the most reliable estimates for general word probabilities. Likewise, a model built from a topic-specific subset of the training data should have the most reliable estimates for on-topic words. It may not be ideal to predict the probability of a word by combining estimates from language models built for different purposes. Our novel nonlinear interpolation scheme uses a general model and a topic-specific model, and a three-way division of the vocabulary into general, on-topic and off-topic subsets. The general and off-topic word probabilities are provided by the general model, and the on-topic word probabilities are provided by the topic model. The off-topic word probabilities are scaled downward to better match their total probability in the topic data.

Other methods of topic adaptation have been explored that do not involve the interpolation of models. Examples of these techniques, such as unnormalized exponential models, dynamic marginals, and topic coherence, can be found in [2, 6, 9].

2. TOPIC ADAPTATION

To adapt a language model to topic, the articles in the training corpus are clustered into possibly overlapping topical subsets using either manually-assigned topic labels, as in our work, or automatic clustering techniques, as in [3, 4, 5, 7]. Each cluster is considered representative of a topic, and only contains articles related to that topic.

We perform topic adaptation in the context of speech recognition.

A first-pass transcription hypothesis for each article in a test set is generated by a speech recognizer using a general language model trained on the entire training corpus. A naive Bayes classifier uses that hypothesis to identify the topic clusters that are most similar to the article. In particular, we select the topics t with the highest posterior probabilities $p(t|D)$ given the hypothesis data D, where we take

$$p(t \mid D) \propto p(t)p(D \mid t) = p(t) \prod_{w_i \in D} p_s(w_i \mid t) \qquad (2)$$

The probability $p(D|t)$ of each topic t generating the hypothesis is calculated using a smoothed estimate of the topic unigram distribution $p_s(w_i|t)$. The smoothed distribution is an interpolation of the unigram distribution $p(w_i|t)$ estimated from the text in the topic cluster and the general unigram distribution $p(w_i)$ estimated from the entire training corpus, *i.e.*,

$$p_s(w_i \mid t) = (1 - \lambda)p(w_i \mid t) + \lambda p(w_i) \qquad (3)$$

The interpolation parameter λ was empirically chosen to be 0.25. The topic priors $p(t)$ are computed from the topic document frequencies. For each article in the test set, a topic specific language model is built by combining the text from the five most similar clusters chosen by the naive Bayes classifier.

2.1. General vs. Topical Words

A vocabulary is chosen consisting of the most frequent words from the entire training corpus. The vocabulary is first divided into two sets: the set of general words and the set of topical words. This division is made independent of topic, so that one division of the vocabulary can be used for any set of topics that are selected for a test set article. Two ways to make this division are presented: Hotelling's T^2 test and Kullback-Leibler distance.

Hotelling's T^2 test Hotelling's T^2 test is used to test whether the mean vectors of two independent random samples of observations on some multidimensional variate are sampled from the same distribution. This test is used as a test of generality vs. topicality for a particular word w by dividing all training set articles into two groups — those that contain w and those that do not.

For each group of articles, a mean vector is constructed containing as many elements as topics, where each element of the vector is the number of articles belonging to that topic in the group divided by the total number of articles in the group.

The Hotelling T^2 statistic is defined as

$$T^2 = n_1 n_2 (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{C}^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)/(n_1 + n_2) \qquad (4)$$

where n_1 and n_2 are the number of articles in each group, $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ are the mean vectors of each group, and $\mathbf{C}$ is the pooled covariance matrix. This statistic tells us whether the distribution of articles across topics depends significantly on the presence of the word w in those articles. A large value for the T^2 statistic is evidence that the mean vectors are significantly different for the two groups of articles, indicating that the word w that determined the article group split is a topical word.

Kullback-Leibler distance The Kullback-Leibler distance is measured between $p(t)$, the *a priori* topic distribution, and $p(t|w)$, the distribution across topics given the word w:

$$D(p(t) \parallel p(t \mid w)) = \sum_{t \in T} p(t) \log (p(t)/p(t \mid w)) \qquad (5)$$

The *a priori* topic distribution $p(t)$ is determined by dividing the number of articles in a topic by the total number of articles. The distribution $p(t|w)$ is calculated by dividing the number of articles in topic t containing word w by the total number of articles containing word w. General words are expected to correspond to small distance values, since knowing these words should not change the topic distribution much. Topical words are expected to have large values, since they would skew $p(t|w)$ away from $p(t)$ by providing strong evidence for certain topics.

2.2. On-Topic vs. Off-Topic Words

Once the vocabulary has been divided into general and topical words, the set of topical words is further divided into a set of on-topic and off-topic words relative to the five most similar topics chosen for each test set article by the naive Bayes classifier. Two different ways to make this split are considered: the χ^2 test and average mutual information.

χ^2 Test The χ^2 test tells us whether a word w occurs significantly more times in topic t than would be expected in general. For each word in a given topic, the following is computed:

$$(O_w - E_w)^2 / E_w$$

where O_w is the observed number of articles containing word w in the current topic and E_w is the expected number of articles containing word w in the current topic. E_w is calculated by multiplying the number of articles in the current topic by the proportion of articles containing word w in the entire training corpus. A χ^2 value is calculated for all words for which $O_w > E_w$, and words with above-threshold values are considered on-topic.

Average Mutual Information The average mutual information between a word and a topic is:

$$\begin{aligned} I(w;t) \quad = \quad & p(w,t) \log \frac{p(t \mid w)}{p(t)} + p(w,\overline{t}) \log \frac{p(\overline{t} \mid w)}{p(\overline{t})} \\ & + p(\overline{w},t) \log \frac{p(t \mid \overline{w})}{p(t)} + p(\overline{w},\overline{t}) \log \frac{p(\overline{t} \mid \overline{w})}{p(\overline{t})} \end{aligned}$$

where $p(w,t)$ is the proportion of articles that are in topic t and contain the word w. Average mutual information measures the amount of information that the presence of a word in an article provides about whether that article is labeled with the given topic. This value is calculated for every word relative to each topic. Words with a high average mutual information for a specific topic are considered on-topic, whereas words with a low value are off-topic.

2.3. Nonlinear Interpolation

Once there is a general and a topic-specific language model for a test article and a three-way division of the vocabulary into general, on-topic and off-topic words, the two models can be interpolated based on the three word lists. Words in the general word

list V_G are predicted from the general language model p_g, words from the on-topic word list V_{ON} are predicted from the topic-specific language model p_t, and words from the off-topic word list V_{OFF} are predicted from the general language model:

$$\begin{aligned} w \in V_G: &\quad p(w \mid h) = p_g(w \mid h) \\ w \in V_{\text{ON}}: &\quad p(w \mid h) = \lambda_{\text{ON}}(h) p_t(w \mid h) \\ w \in V_{\text{OFF}}: &\quad p(w \mid h) = \lambda_{\text{OFF}}(h) p_g(w \mid h) \end{aligned}$$

The scale factors $\lambda_{\text{ON}}(h)$ and $\lambda_{\text{OFF}}(h)$ are calculated so that the general words occupy as much probability mass in the adapted model as they do in the general model. The on-topic and off-topic words then split the remaining probability mass in the same proportion as they do in the topic-specific model. As a result, the on-topic words generally occupy more probability mass in the adapted model than in the general model (they have been boosted), and the off-topic words occupy less probability mass (they have been suppressed.) The scale factors are computed as follows:

$$\begin{aligned} m_{g\_V_G}(h) &= \sum_{w \in V_G} p_g(w \mid h) \\ m_{t\_V_G}(h) &= \sum_{w \in V_G} p_t(w \mid h) \\ m_{g\_V_{\text{ON}}}(h) &= \sum_{w \in V_{\text{ON}}} p_g(w \mid h) \\ m_{t\_V_{\text{ON}}}(h) &= \sum_{w \in V_{\text{ON}}} p_t(w \mid h) \\ \lambda_{\text{ON}}(h) &= \frac{1 - m_{g\_V_G}(h)}{1 - m_{t\_V_G}(h)} \\ \lambda_{\text{OFF}}(h) &= \frac{(1 - m_{g\_V_G}(h))(1 - m_{t\_V_G}(h) - m_{t\_V_{\text{ON}}}(h))}{(1 - m_{t\_V_G}(h))(1 - m_{g\_V_G}(h) - m_{g\_V_{\text{ON}}}(h))} \end{aligned}$$

3. EXPERIMENTS

We evaluated our topic adaptation algorithm on a Broadcast News training and test set. The training data consists of 130M words and 88k articles. Each article is accompanied by a set of topic labels that describe the article's topic[1]. The corpus was split into topic clusters by assigning each topic label to a cluster. The text for each article was assigned to the clusters of the article's labels. A total of 5883 clusters were available for topic adaptation. The most frequent 51k words from the training corpus were selected as the vocabulary, and a general trigram language model was built with the CMU language modeling toolkit [8].

Hotelling's T^2 test and the Kullback-Leibler distance were used to rank the words in the vocabulary from general to topical. The Kullback-Leibler distance was computed using a topic distribution across all 5883 topic clusters, but for the T^2 statistic (which involves a matrix inversion), the 5883 clusters were mapped down to 50 clusters using an agglomerative clustering technique as described in [10]. Thresholds were set on these two ranked lists, dividing the words into general and topical sets. Additionally, a 595-word stopword list derived from the SMART system stopword list[2] was used as the general word list.

[1] The topic labels were provided by the transcribers of the training text.

[2] Available at ftp://ftp.cs.cornell.edu/pub/smart/smart.11.0.tar.Z

The test set consists of 57 stories from the Hub-4 1996 development set. For each article, a naive Bayes classifier was used to select the most similar five topic clusters, and the text from these clusters was combined to build a topic-specific language model. The χ^2 and average mutual information methods were used to create ranked topical word lists for each of the 5883 topic clusters. An on-topic word list was generated for each test article by traversing the topical word lists in descending order of score for each of the five selected topic clusters, until k words from the general word list were encountered, where we considered $k = 1$ and $k = 10$. The selected words from the five lists were combined to make the on-topic word list. All words from the vocabulary that were not assigned to either the general or on-topic word lists were assigned to the off-topic word list. The word lists were used to interpolate the general and topic-specific models for each of the 57 articles.

Table 1 shows the perplexity values obtained on the reference transcripts of the test set, using the general language model only, the topic-specific language models only, linear interpolation of the general and topic-specific language model for each story, and the interpolated language models for various selection configurations of the general, on-topic and off-topic word lists. MI indicates that the topic lists were derived using the average mutual information measure. The -1 and -10 designations indicate that on-topic words were collected from each of the five topical word lists until either 1 or 10 general words were encountered. KL and $stop$ correspond to the general word lists derived from the Kullback-Leibler measure and the stopword list, and the numbers in parentheses are the number of words in the general word list. Linear interpolation of the general and topic-specific language models used two-way cross-validation to choose interpolation weights for each test story.

General LM PP: 189			
Story LM PP: 236			
Linear Interpolation PP: 174			
	General word rankings		
Topic word rankings	$T^2(595)$	T^2 (1736)	stop (595)
χ^2-1	187	184	181
χ^2-10	189	186	181
	KL (595)	KL (2000)	stop (595)
MI-1	182	184	182
MI-10	181	182	183

Table 1: Perplexity results using various configurations on general, on-topic and off-topic word lists.

Using the general language model alone results in a perplexity value of 189. The best nonlinear interpolation result was 181, when the stopword list was used with the χ^2 lists, or when the Kullback-Leibler general list was used with the average mutual information topic list. Linear interpolation achieves a perplexity value of 174.

Table 2 shows word error rate (WER) results from rescoring N-best lists generated by the Sphinx-3 decoder for the three nonlinear interpolation configurations that produced the lowest perplexity values. The WER of the hypothesis transcriptions (Hyp)

used for topic detection is 40.2%. The lowest achievable N-best rescoring WER (Lowest), found by using the reference transcripts to choose the N-best hypotheses with the lowest error, was 34.6%. Using the general language model to rescore the N-best lists results in a WER of 40.1%. The interpolated language models result in a WER of 39.8% in all three cases.

Hyp	40.2%
Lowest	34.6%
General LM	40.1%
χ^2-1, stop-595	39.8%
χ^2-10, stop-595	39.8%
MI-10, KL-595	39.8%

Table 2: Word error rate results from N-best rescoring using best three configurations of general, on-topic and off-topic word lists.

4. DISCUSSION

Although nonlinear interpolation does result in a decrease in perplexity (4%) and WER over using a general language model alone, the magnitude of the decrease is not as great as that obtained with linear interpolation (8% decrease in perplexity.) We were surprised that nonlinear interpolation did not perform better, and began examining the MI-10, KL-595 configuration more closely in order to determine the reason for the lack of perplexity improvement. On average, 264 words were chosen as on-topic from the average mutual information lists for each of the 57 test articles. The test set consists of 23,082 in-vocabulary word tokens: 15,963 are general, 2,049 are on-topic, and 5,070 are off-topic. The perplexity values for predicting the word class (general, on-topic, or off-topic) given the history, and then predicting the word given the class for the general, topic-specific and adapted models are shown in Table 3. The adapted model does slightly better at predicting the class than the general and topic-specific models, which shows that the scaling of the on-topic and off-topic words has helped the adapted model. The general model does better than the topic-specific models at predicting the general and off-topic words, as hoped. However, the topic-specific models do no better at predicting the on-topic words than the general model. Ideally, the topic-specific models would provide a much lower perplexity for the on-topic words than the general model, which is not the case for this adaptation configuration. We are continuing to investigate the reasons for the higher than expected perplexity from the topic-specific models by considering the selection of data for these models and the choice of on-topic words. Further analysis and results will be reported at the conference and at http://www.cs.cmu.edu/People/kseymore/icslp98.html.

5. ACKNOWLEDGMENTS

We would like to thank Larry Wasserman for his help in developing several of the ideas in this paper. This material is based upon work supported under a National Science Foundation Graduate Research Fellowship and by the Department of the Navy, Naval Research Laboratory, under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government or the National Science Foundation.

	Class	General	On-topic	Off-topic
General LM	1.99	39	53	1,945
Topic LM	1.93	48	54	3,279
Adapted LM	1.91	39	54	1,945

Table 3: Perplexity results for $P(class)$ and $P(word|class)$ where *class* $\in$ {*general, on-topic, off-topic*} for the general, topic-specific and adapted models for configuration MI-10, KL-595.

6. REFERENCES

1. S. Chen, A. Kalai, A. Blum, and R. Rosenfeld. On-line algorithms for combining language models. 1998. In preparation.
2. S. Chen, K. Seymore, and R. Rosenfeld. Topic adaptation for language modeling using unnormalized exponential models. In *Proc. ICASSP-98*, 1998.
3. P. Clarkson and A. Robinson. Language model adaptation using mixtures and an exponentially decaying cache. In *Proc. ICASSP-97*, pages 799–802, 1997.
4. R. Iyer and M. Ostendorf. Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models. In *Proc. ICSLP*, pages 236–239, 1996.
5. R. Kneser and J. Peters. Semantic clustering for adaptive language modeling. In *Proc. ICASSP-97*, volume 2, pages 779–782, 1997.
6. R. Kneser, J. Peters, and D. Klakow. Language model adaptation using dynamic marginals. In *Proc. Eurospeech '97*, 1997.
7. S. Martin, J. Liermann, and H. Ney. Adaptive topic-dependent language modelling using word-based varigrams. In *Proc. Eurospeech '97*, 1997.
8. R. Rosenfeld. The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation. In *Proceedings of the Spoken Language Systems Technology Workshop*, pages 47–50, Austin, Texas, January 1995.
9. S. Sekine and R. Grishman. NYU language modeling experiments for the 1995 CSR evaluation. In *Proc. of the ARPA Spoken Language Systems Technology Workshop*, 1995.
10. K. Seymore and R. Rosenfeld. Large-scale topic detection and language model adaptation. Technical report, Carnegie Mellon University, June 1997.
11. K. Seymore and R. Rosenfeld. Using story topics for language model adaptation. In *Proc. Eurospeech'97*, 1997.