Chapter 4

Topic Dependent Language Models

In this chapter, we explore topic dependencies between words and documents and improve the performance of language models by exploiting topic dependencies. We cluster the training data into topics, extract topic-dependent salient features from each topic and build maximum entropy models with both topic constraints and regular N-gram constraints. The training issues have been stated in Chapter 3. We will provide experimental results with detailed analysis based on Switchboard and Broadcast News. We also compare the maximum entropy method with other approaches reported in the literature.

4.1 Introduction

N-gram language models treat words in sentences simply as abstract symbols without considering their semantic roles in the language and their syntactic roles in sentences. The speech recognizer with a regular N-gram model may not be able to choose the correct word from two words with close pronunciations and similar collocational frequencies. For example, gas and guys after the words lot of are difficult to recognize in Switchboard because they are acoustically similar and have the same relative frequencies: \( f(gas|lot\ of) \) and \( f(guys|lot\ of) \) both equal 0.0006.

Humans benefit from both semantic and syntactic knowledge in understanding natural language. If people hear the potentially confusing utterances lot of gas/guys
when discussing buying a car, they may tend to choose “gas” instead of “guys” since the former one is relevant to the topic of discussion while the latter one is less so. Actually the relative frequency of $f(\text{gas}|\text{lot of})$ is boosted 15 fold to 0.009 in the topic \textit{buying a car} in Switchboard while $f(\text{guys}|\text{lot of})$ drops nearly to zero.

In general, the use of words in sentences depends on the topic of discussion. This relationship has been successfully used to discern the topic of a document in information retrieval (IR). In language modeling, an accurate estimation of the probability of these topic relevant words may be obtained via determination of the semantic content of the document, \emph{e.g.}, by topic detection.

To understand the influence of topics on word frequencies quantitatively, we show a concrete example from the Switchboard corpus, a human to human telephone speech database on 70 assigned topics such as \textit{education}, \textit{sports}, \textit{child care} and \textit{automobile}. Each topic contains several hundred to several thousand utterances, and thousands to tens of thousands of words. The whole corpus comprises about 260,000 sentences, amounting to about 3 million words. A vocabulary of about 22,000 words is chosen to cover all words in the training and test data. The frequency of many content-bearing words in a specific topic varies significantly from that in the whole corpus. Table 4.1 lists the relative frequency of some words in the topic \textit{clothing} and that in the whole Switchboard training corpus. The former (in Column 2) is at least one order of magnitude greater than the latter (in Column 3).

A straightforward way of adding the topic information in language models is to classify the training data according to the topic and estimate a topic-specific N-gram model $p_t(w_i|w_{i-N+1}, \ldots, w_{i-1})$ for each topic $t$ from the training data of this topic. Unfortunately, there exists a severe data sparseness problem. In Switchboard, no topics have more than one hundred thousand training samples (words), which is in the same order of magnitude as the vocabulary size. The estimate of a trigram probability $p_t(w_i|w_{i-2}, w_{i-1})$ (with 8 trillion parameters) will be very unreliable based on such a small training set. Of course, such a probability estimate can be interpolated with or backed-off to an overall trigram probability (or the topic-dependent bigram probability) as in Iyer & Ostendorf (1996) and Florian & Yarowsky (1999). However, this implementation is not space efficient since each topic must have its own language
<table>
<thead>
<tr>
<th>Word</th>
<th>Freq in Clothing</th>
<th>Freq in whole Corpus</th>
<th>log difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPEARANCE</td>
<td>0.00045181</td>
<td>1.57886e-05</td>
<td>1.4566</td>
</tr>
<tr>
<td>ATTIRE</td>
<td>0.00045181</td>
<td>3.94714e-06</td>
<td>2.0586</td>
</tr>
<tr>
<td>ATTORNEYS</td>
<td>0.00060241</td>
<td>2.76300e-05</td>
<td>1.3385</td>
</tr>
<tr>
<td>AVON</td>
<td>0.00030121</td>
<td>7.89428e-06</td>
<td>1.5815</td>
</tr>
<tr>
<td>BACKLESS</td>
<td>0.00030121</td>
<td>2.63143e-06</td>
<td>2.0587</td>
</tr>
<tr>
<td>BAKERY</td>
<td>0.00060241</td>
<td>1.05257e-05</td>
<td>1.7576</td>
</tr>
<tr>
<td>BLOUSE</td>
<td>0.00060241</td>
<td>1.05257e-05</td>
<td>1.7576</td>
</tr>
<tr>
<td>BLOUSES</td>
<td>0.00090361</td>
<td>1.05257e-05</td>
<td>1.9337</td>
</tr>
<tr>
<td>BOOTS</td>
<td>0.00090361</td>
<td>1.97357e-05</td>
<td>1.6607</td>
</tr>
<tr>
<td>BAGGY</td>
<td>0.00030121</td>
<td>3.94714e-06</td>
<td>1.8847</td>
</tr>
<tr>
<td>BOOTS</td>
<td>0.00090361</td>
<td>1.97357e-05</td>
<td>1.6626</td>
</tr>
</tbody>
</table>

Table 4.1: Sensitive words for the topic “CLOTHES”.

Instead of using either the interpolation or the back-off method, we build a maximum entropy model to capture the topic-dependencies. In our method, unigram frequencies in a specific topic $t$ are treated as *topic-dependent* salient features, just as overall N-gram frequencies are *topic-independent* salient features. Topic-dependent models $p(w_i|w_{i-N+1}, \ldots, w_{i-1}, t)$ are constructed in which the probability of a word $w_i$ depends not only on the N-1 preceding words, $w_{i-N+1}, \ldots, w_{i-1}$, but also on the topic $t$ of the history. The maximum entropy method is used to select the “smoothest” statistical model from these models.

### 4.2 Exploiting Topic Dependence

To obtain the topic-sensitive features, we first classify the training data by topic and for each topic extract words whose frequencies within the topic deviate significantly from their overall distribution. Robust topic detection algorithms used in information retrieval can be applied to the clustering of training data and topic assignment for the test data.
4.2.1 Topic Classification

We construct topic-dependent language models for both Switchboard and Broadcast News (BN). The Switchboard corpus has been introduced in Section 4.1. Here, we briefly introduce the Broadcast News corpus. The Broadcast News corpus consists of news dispatches (stories) broadcast between 1994 and 1996. The training set contains about 125,000 stories amounting to about 130 million words (including the special end-of-sentence token). We treat each conversion side in Switchboard or story in Broadcast News as an individual document that is the basic granule for topic clustering. For each document, we create a term frequency/inverse document frequency (TF-IDF) vector of all words in the vocabulary excluding stop words\footnote{A stop word is any of a set of function words with low semantic content that will be ignored by the topic classifier.}, where the TF-IDF unigram frequency \( x_i \) for the word \( w_i \) is defined as

\[ x_i = f(w_i) \cdot \log \frac{D}{D_{w_i}} \]

in which \( D \) and \( D_{w_i} \) are the total number of documents and the number of documents containing \( w_i \), respectively. The factor \( \log \frac{D}{D_{w_i}} \) boosts the frequency of topic-specific words while lessening the weight for those words appearing in most of the topics.

We let \( X = x_1, \cdots, x_{|V|} \) be the TF-IDF vector of the document where \( |V| \) is the vocabulary size and \( Y \) be the similar vector for the centroid. We then define the smoothed cosine distance following Florian & Yarowsky (1999) as

\[ D(X, Y) = 1 - \frac{X \cdot Y + \epsilon \sum_{x_i > 0} x_i + \epsilon \sum_{y_i > 0} y_i + \epsilon^2}{(\| X \| + \epsilon) \cdot (\| Y \| + \epsilon)} \]  \hspace{1cm} (4.1)\]

where \( \| \cdot \| \) denotes the \( L_2 \) norm for the parameter and \( \epsilon \) is a small positive real number\footnote{\( \epsilon = 0.2 \) in our experiments.} less than 1.

We use the following standard K-means clustering method to classify conversations into different topics.

**Algorithm (K-means Clustering)**

**Initial Step:**
Determine a set of \( K \) centroids according to pre-assigned topic labels of
each document. 
Let \( C^0(X_i) \) be the initial class for the document \( X_i, i = 1, 2 \cdots, D \), and set
\[
Y_j^0 = \sum_{i: C^0(X_i) = j} X_i \quad \text{for} \quad j = 1, 2, \cdots, K.
\]

In Switchboard, the initial cluster assignments are derived from the 70 manually assigned topics of the conversations, whereas in Broadcast News, initial topic tags are assigned by the bottom-up hierarchical clustering algorithms provided by Florian & Yarowsky (1999).

**Iteration Steps:**
- Assign each document to its nearest centroid according to cosine distance.
- Let \( C^n(X_i) \) be the new centroid for \( X_i, i = 1, 2, \cdots, D \), where
\[
C^n(X_i) = \arg \min_{j=1}^{K} D(X_i, Y_j^{n-1}).
\]

Note that \( D(X_i, Y_j^{n-1}) = \infty \) if \( Y_j \) is empty.
- Recalculate the centroids based on the new clustering of dialogues.
\[
Y_j^n = \sum_{i: C^n(X_i) = j} X_i \quad \text{for} \quad j = 1, 2, \cdots, K.
\]

Repeat the iteration steps until the centroids are “stable” \(^3\). 

Sixty-seven of seventy initial clusters are obtained for the Switchboard. Only 3% of the conversation sides out of 2200 are assigned different topics from their initial ones after re-clustering. For the Broadcast News, 100 clusters are generated and about 20% of stories change topics after the re-clustering.

### 4.2.2 Selection of Topic Sensitive Words

We design a non-singleton word \( w \) as a topic-related word if its relative frequency \( f_t(w) \) in topic \( t \) differs significantly from its relative frequency \( f(w) \) in all training

\(^3\)If a document \( X_i \) of topic \( Y_j \) is closer to another cluster \( Y_{j'} \) but the difference of cosine distances is not greater than a threshold (very small), it does not switch to \( Y_{j'} \) but stays in the original \( Y_j \).
data, i.e.,

$$|D(f_i(w)||f(w))| = \left| f_i(w) \cdot \log \frac{f_i(w)}{f(w)} \right| \geq T_i$$

where $D$ is the I-divergence and $T$ is a threshold. In Switchboard, we find 8459 different topic-sensitive words (40% of the vocabulary) and 15442 topic-word combinations ($t,w$ pairs). Each topic has 231 sensitive words on average, and each topic-sensitive word belongs to 1.8 topics on average (and 8 maximum). In Broadcast News, 95% of words (59989 out of 63032) in the vocabulary are topic-sensitive words. Each topic has about 2,500 topic-sensitive words on average and each words belongs to 5 topics on average (with the maximum number being 15). There are a total 249295 topic-word combinations.

### 4.2.3 Formulation of the ME Topic Model

We use the long-range history $w_1, \ldots, w_{i-1}$ to assign a topic $t_i = t(w_1, \ldots, w_{i-1})$ to a test utterance. The sufficient statistic of the history is thus the triple $\phi(h_i) = [w_{i-1}, w_{i-2}, t_i]$, and we set

$$p(w_i|w_1, \ldots, w_{i-1}) \approx p(w_i|w_{i-1}, w_{i-2}, t_i). \quad (4.2)$$

Of course, not every word in the vocabulary will have strong dependence on the topic. Estimating a separate conditional pmf for each $[w_{i-1}, w_{i-2}, t_i]$ fragments the training data and may result in poor estimates for these words. In additional, topic-related words may not appear in every word-context $[w_{i-1}, w_{i-2}]$. We therefore seek a model which, in addition to topic-independent N-gram constraints,

$$\sum_{t_i} p(w_i|w_{i-1}, w_{i-2}, t_i)p(w_{i-2}, w_{i-1}, t_i) = \frac{\#[w_{i-2}, w_{i-1}, w_i]}{\#\text{[training data]}}^4, \quad (4.3)$$

$$\sum_{t_i, w_{i-2}} p(w_i|w_{i-1}, w_{i-2}, t_i)p(w_{i-2}, w_{i-1}, t_i) = \frac{\#[w_{i-1}, w_i]}{\#\text{[training data]}}, \quad (4.4)$$

$$\sum_{t_i, w_{i-2}, w_{i-1}} p(w_i|w_{i-1}, w_{i-2}, t_i)p(w_{i-2}, w_{i-1}, t_i) = \frac{\#[w_i]}{\#\text{[training data]}}, \quad (4.5)$$

\(^4\)We choose $T_i = \frac{3}{C_t}$ where $C_t$ is the number of words in document $t$. 

meets topic-dependent marginal constraints

\[
\sum_{w_{i-1}, w_{i-2}} p(w_i | w_{i-1}, w_{i-2}, t_i) p(w_{i-1}, w_{i-2}, t_i) = \frac{\#[t_i, w_i]}{\#[\text{training data}]}.
\] (4.6)

Note that these marginal probabilities are much more reliably estimated from counts \#[\cdot] of the corresponding events observed in the corpus than are the conditional probabilities in (4.2). For example, only 2% of 4-tuples \((w_{i-2}, w_{i-1}, w_i, t_i)\) in the Switchboard test data occur in the training data. However, for 63% of the test data, either \((w_{i-2}, w_{i-1}, w_i)\) or \((t_i, w_i)\) does not appear in the training set.

Unreliable marginal probabilities, e.g., those based on one or two observations, may be completely left out of the model’s requirements. In addition to leaving out some constraints, the relative frequency estimates of the marginal probabilities on the right-hand sides of equations (4.3)-(4.6) may be replaced by their corresponding Good-Turing (Good, 1953) estimates. Finally, since our primary interest is in the conditional model \(p(w_i|w_{i-1}, w_{i-2}, t_i)\), empirically observed relative frequencies \(\hat{p}(w_{i-1}, w_{i-2}, t_i)\) may be substituted for the model-based probabilities \(p(w_{i-1}, w_{i-2}, t_i)\) in equations (4.3)-(4.6).

Linear constraints of the form described in (4.3)-(4.5) define a family of pmfs, and we choose the model in this family that has the highest entropy, corresponding qualitatively to the least additional assumptions on (or maximal smoothness of) the model. As explained in Chapter 2, the ME model has an exponential form, with one parameter \(\alpha\) corresponding to each linear constraint placed on the model:

\[
p(w_i | w_{i-1}, w_{i-2}, t_i) = \frac{g(w_i) \cdot g(w_{i-1}, w_i) \cdot g(w_{i-2}, w_{i-1}, w_i) \cdot \alpha^{(t_i, w_i)}}{z(w_{i-1}, w_{i-2}, t_i)}
\] (4.7)

where \(z(w_{i-1}, w_{i-2}, t_i)\) is a suitable normalization constant. The first three numerator terms correspond to standard N-gram constraints, whereas the fourth one is a topic-unigram parameter determined by word-frequencies in particular topics.

### 4.2.4 Topic Assignment to the Test Utterance

To use a topic-dependent model for rescoring, a topic must be assigned to test utterances automatically. Two issues arise when a topic-dependent language model
for speech recognition is used. First, since the actual spoken words (or reference transcriptions) are not available for topic assignment, topic assignment must be based on recognition results. We study the impact of recognition errors on this process. Second, the topic of a document may change as the document progresses. We examine whether a topic should be assigned to an entire test conversation or to each utterance. If the topic assignment is on the utterance level, it can be based on

(i) that utterance only,
(ii) that utterance and a few preceding utterances, or
(iii) that utterance, a few preceding and following utterances.

The results will be presented in Section 4.3.2.

To detect topics automatically, the weighted TF-IDF vectors based on real hypotheses of the test utterances (instead of truth) are generated for each document or each utterance. The closest centroid is chosen for each document or each utterance according to the cosine similarity between vectors of the test data and the centroids.

4.3 Switchboard Results

Most of the experiments are based on the Switchboard task. We use 1,100 conversations amounting to 2.1 million words to train the language model. The acoustic models used for recognition are state-clustered cross-word triphone HMMs with 12 Gaussian mixture output densities, trained with MF-PLP acoustic features from about 60 hours of speech data. The test set (WS97 dev-test set) has a little over 2 hours of speech in the form of 19 conversations containing about 2400 utterances amounting to 18,000 words.

For every test utterance, a list of the 100-best hypotheses is generated by an HTK-based recognizer (Young et al., 1995) using a back-off trigram language model. The trigram model has about 500K constraints and the topic-dependent model contains 15K topic constraints besides N-grams. The recognition word error rate (WER) for rescoring these hypotheses and the average perplexity of the transcriptions of the test set are reported here.

5The 10 best-hypotheses are used in this research
4.3.1 Baseline Results

Table 4.2 shows the performance of standard Katz (1987) backoff trigram models built by SRI LM toolkits (Stolcke, 1996) and an ME model with only Ngram constraints. The minimum count for a bigram to be included in a model is indicated by B, and that for a trigram by T. All counts of less than 8 are replaced by their corresponding Good-Turing (Good, 1953) estimates here and also in the remainder of the dissertation. It should be noted that singleton trigrams have no constraints on their probability in the ME model.

The back-off model with no bigram cutoff (i.e., the bigram cutoff of 1) and the trigram cutoff of 2 performs best among all back-off trigram models. The ME model performs only slightly (but significantly⁶) better than the corresponding back-off model when only N-gram constraints are specified. Thus any further improvement of the language model by adding long-range dependence may be credited to adding those new features to N-gram constraints rather than to any inherent advantage of the ME method.

<table>
<thead>
<tr>
<th>Model (N-gram cutoffs)</th>
<th>Perplexity</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-off (no cutoffs)</td>
<td>79.2</td>
<td>38.7%</td>
</tr>
<tr>
<td>Backoff ($B \geq 1$, $T \geq 2$)</td>
<td>78.8</td>
<td>38.5%</td>
</tr>
<tr>
<td>ME Trigram</td>
<td>78.9</td>
<td>38.3%</td>
</tr>
</tbody>
</table>

Table 4.2: Perplexity and WER of back-off trigram models and an ME model with the same constraints

4.3.2 Topic Assignment During Testing

The topic model (4.7) is trained next. The recognition performance depends on the detection of the topic. A hard decision is made by assigning the closest matching topic in the results presented here, although the formalism extends easily to soft topic decisions. We first fix the topic for the whole conversation and test the performance of our language models. Next, we let the topic shift inside a conversation.

⁶Based on the sign-test introduced in Appendix A.
**Topic Assignment Based on Conversations**

We investigate four options for this assignment:

(i) manual assignment of topics to the conversation,
(ii) automatic topic assignment based on the reference transcriptions, \(^7\)
(iii) automatic topic assignment based on the 10-best hypotheses generated by the first recognition pass, and
(iv) oracle assignment to minimize perplexity of the test set.

<table>
<thead>
<tr>
<th>Source of Text for Topic Classification</th>
<th>Perplexity</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (Baseline)</td>
<td>78.8</td>
<td>38.5%</td>
</tr>
<tr>
<td>Manual Assignment</td>
<td>73.1</td>
<td>37.8%</td>
</tr>
<tr>
<td>Ref. Transcriptions</td>
<td>73.8</td>
<td>37.8%</td>
</tr>
<tr>
<td>10-Best Hypotheses</td>
<td>74.4</td>
<td>37.9%</td>
</tr>
<tr>
<td>Oracle (optimal)</td>
<td>72.5</td>
<td>37.7%</td>
</tr>
</tbody>
</table>

Table 4.3: Topic assignment based on erroneous recognizer hypotheses causes little degradation in performance.

Manual and automatic topic assignments are listed in Tables 4.4 and 4.5, in which any automatically assigned topic identical to the manual one is replaced by a *. It is apparent that 34 out of 38 conversation sides are assigned to the same topics (Column 3) as the manual tags (Column 2) when reference transcripts are used. The rest 4 conversation sides are assigned topics close to the manual ones. For example, conversation #2461A is assigned to the topic *choosing a college*, whereas the manual is *public education*. When the 10-best hypotheses are used instead of truth, 32 conversations are assigned to the same topics (Column 4) as the manual ones. These results indicate that the quality of automatic topic classification is quite good.

The results, presented in Table 4.3, clearly indicate that even with a WER of over 38%, there is only a small loss in perplexity and a negligible loss in WER when the

\(^7\)The null topic, which defaults to a topic-independent baseline model, is available to the topic classifier as one of the choices.
<table>
<thead>
<tr>
<th>Conv. Id</th>
<th>Manual</th>
<th>Ref.</th>
<th>N-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>2121A</td>
<td>air pollution</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2121B</td>
<td>air pollution</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2131A</td>
<td>music</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2131B</td>
<td>music</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2151A</td>
<td>universal public service</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2151B</td>
<td>universal public service</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2229A</td>
<td>exercise and fitness</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2229B</td>
<td>exercise and fitness</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2335A</td>
<td>crime</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2335B</td>
<td>crime</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2434A</td>
<td>gun control</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2434B</td>
<td>gun control</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2441A</td>
<td>universal public service</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2441B</td>
<td>universal public service</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2461A</td>
<td>public education</td>
<td>choosing a college</td>
<td>choosing a college</td>
</tr>
<tr>
<td>2461B</td>
<td>public education</td>
<td>*</td>
<td>choosing a college</td>
</tr>
<tr>
<td>2503A</td>
<td>pets</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2503B</td>
<td>pets</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 4.4: Topic assignments of test conversations using different methods.

topic assignment is based on recognizer hypotheses instead of the correct transcriptions. Comparisons with the oracle result indicate that there is little room for further improvement.

**Topic Assignment Based on Utterances**

A more precise topic assignment may be based on the utterance level. The topic of an utterance can be determined by

(i) itself, or
(ii) several preceding utterances and itself, or
(iii) several preceding and following utterances and itself.

Each utterance used for topic detection is assigned a positive weight $\omega$. The maximum weight is assigned to the current utterance to emphasize its importance. The weight reduces linearly with an increase in the distance between that utterance and the
<table>
<thead>
<tr>
<th>Conv. Id</th>
<th>Manual</th>
<th>Ref.</th>
<th>N-best</th>
</tr>
</thead>
<tbody>
<tr>
<td>2632A</td>
<td>public education</td>
<td>choosing a college</td>
<td>choosing a college</td>
</tr>
<tr>
<td>2632B</td>
<td>public education</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2724A</td>
<td>exercise and fitness</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2724B</td>
<td>exercise and fitness</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2752A</td>
<td>air pollution</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2752B</td>
<td>air pollution</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2753A</td>
<td>latin america</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2753B</td>
<td>latin america</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2836A</td>
<td>crime</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2836B</td>
<td>crime</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2838A</td>
<td>gun control</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2838B</td>
<td>gun control</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3528A</td>
<td>space exploration</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3528B</td>
<td>space exploration</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3756A</td>
<td>painting</td>
<td>home repair</td>
<td>home repair</td>
</tr>
<tr>
<td>3756B</td>
<td>painting</td>
<td>home repair</td>
<td>home repair</td>
</tr>
<tr>
<td>3942A</td>
<td>ethics in government</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3942B</td>
<td>ethics in government</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3994A</td>
<td>music</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>3994B</td>
<td>music</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 4.5: Topic assignments of test conversations using different methods (Cont).

In formal statements, the weight \( \omega(u) \) of an utterance \( u \) is assigned as

\[
\omega(u) = D - d(u)
\]

where \( D^8 \) is the window size and \( d(u) \) is the distance between \( u \) and the current utterance.

The experimental results show that even though a small window may result in lower perplexity than a longer one (Figure 4.1), the lowest word error rate is reached at the point of a moderate window size (Figure 4.2). It also shows that a two-sided window gives only slightly better results than the left-sided one. The difference between using a reference script and the 10-best list is relatively large for a small window since topic-sensitive words may appear only a few times in a small window. If some of them are missed, the topic assignment may not be accurate. However,

---

8 We choose \( D = 5 \).
9 In real time speech recognition, only preceding utterances and the current one are available.
this difference becomes small as the window size increases, since at least some topic-
relevant words may appear enough times in the 10-best list in a window of moderate
size.

![Graph showing perplexities for different window lengths, both left-sided and two-sided](image)

Figure 4.1: Perplexities for different window lengths, both left-sided and two-sided

The best recognition performance in WER (without looking at future utterances)
achieved by assigning a topic to each utterance based on the 10-best hypotheses of
the current and the four preceding utterances. We fix the topic assignment based on
this setup in the rest of our experiments.

We summarize the results of dynamic topic assignment in Table 4.6.

It is worth noting that utterance-level topic assignment of Table 4.6 is slightly more
effective than the conversation level assignment (Table 4.3). Adding topic-dependent
constraints reduces absolute WER by 0.7%, which is very significant\(^\text{10}\), and relative
perplexity by 7%.

\(^{10}\text{With a P-value of } 10^{-5} \text{ in the sign-test, which will be introduced in Appendix A.}\)
Figure 4.2: WER for different (left-sided) window lengths.

To gain insight into improved performance from utterance-level topic assignment, we examine agreement between topics assigned at the two levels. As illustrated in Table 4.7, 8 out of 10 utterances prefer the topic-independent model and are filler utterances, probably serving vital discourse functions (e.g., acknowledgments, back-channel responses). Of the remaining utterances, a majority (60%) are closest to the topic that was assigned at the conversation level. While a large fraction (40%) are closer to a topic other than the one preferred at the conversation level, this is not an equally remarkable result as, in many of these cases, the topic assigned at the conversation-level is a close second or the two topics are similar, e.g., *public education* and *choosing a college*.
### Table 4.6: Dynamic topic assignment for individual utterances based on the current and four preceding utterances.

<table>
<thead>
<tr>
<th>Source of Text for Topic Classification</th>
<th>Perplexity</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (Baseline)</td>
<td>78.8</td>
<td>38.5%</td>
</tr>
<tr>
<td>Ref. Transcriptions</td>
<td>73.3</td>
<td>37.8%</td>
</tr>
<tr>
<td>10-Best Hypotheses</td>
<td>73.5</td>
<td>37.8%</td>
</tr>
</tbody>
</table>

### Table 4.7: Topic dynamics viewed through (dis)agreement of utterance-level and conversation-level topic assignment.

<table>
<thead>
<tr>
<th>Source of Text for Topic Classification</th>
<th>Agreement of Conv. &amp; Utt. Level Topics</th>
<th>Utt. Level Topic When Disagreeing With Conv.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Other Topic</td>
</tr>
<tr>
<td>Ref. Trans.</td>
<td>12.7%</td>
<td>7.1%</td>
</tr>
<tr>
<td>10-Best Hyps.</td>
<td>9.9%</td>
<td>7.0%</td>
</tr>
</tbody>
</table>

### 4.3.3 Analysis of Recognition Performance

Topic-dependent constraints are chosen to manipulate the probabilities of words whose frequency within conversations on a particular topic are dramatically different from their frequency in the whole corpus. These are typically content-bearing words that are interrelated by the topic of the conversation. Thus, we expect improvements from the topic model to be concentrated on such words.

To see if we indeed improve the model where we aim to improve it, the vocabulary is divided into two sets: all words that have topic-conditional unigram constraints for any of the topics, and the others. About 7% of the tokens in the test set have topic-dependent constraints. Errors in recognizing each test token is classified in the same way, *i.e.* insertions or deletions of topic-dependent words as well as a substitution that involves a topic-dependent word counts towards an error in the first category. Table 4.8 shows a breakdown of the results over the set of topic-dependent and topic-independent words for ME models with and without topic-dependent constraints. All perplexity reduction comes from the topic-dependent words. The WER reduction on those topic-dependent words is 2.2%, which is much higher than the overall WER.
reduction.

<table>
<thead>
<tr>
<th>Language Model (N-gram cutoffs)</th>
<th>Topic Words</th>
<th>Non-topic Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perplexity</td>
<td>WER</td>
</tr>
<tr>
<td></td>
<td>Perplexity</td>
<td>WER</td>
</tr>
<tr>
<td>ME (B≥1, T≥ 2)</td>
<td>3795</td>
<td>37.7%</td>
</tr>
<tr>
<td></td>
<td>62.2</td>
<td>38.5%</td>
</tr>
<tr>
<td>ME-Topic (B≥1, T≥ 2)</td>
<td>356</td>
<td>35.3%</td>
</tr>
<tr>
<td></td>
<td>66.6</td>
<td>37.9%</td>
</tr>
</tbody>
</table>

Table 4.8: Analysis of performance gains from the topic-dependent model.

Since the model is designed to benefit topic-dependent words, a fairer comparison may be based on all content-bearing words vs. stop words. Under this partition, about 23% of tokens in the test set are content-bearing and the remainder are stop words. Table 4.9 presents the performance gains analyzed for this partition.

<table>
<thead>
<tr>
<th>Language Model (N-gram cutoffs)</th>
<th>Content Words</th>
<th>Stop Words</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perplexity</td>
<td>WER</td>
</tr>
<tr>
<td></td>
<td>Perplexity</td>
<td>WER</td>
</tr>
<tr>
<td>ME (B≥1, T≥ 2)</td>
<td>8941</td>
<td>42.2%</td>
</tr>
<tr>
<td>ME-Topic (B≥1, T≥ 2)</td>
<td>3923</td>
<td>40.8%</td>
</tr>
</tbody>
</table>

Table 4.9: Performance improvement on content-bearing words.

As expected, Table 4.9 shows that the gain in perplexity comes predominantly from content-bearing words, and the 1.4% improvement in WER on these words is greater than the overall WER improvement. The perplexity for stop words slightly increases, but the WER is also reduced because the correction of any content words will also help to avoid errors in its neighborhood. This may be a more important performance measure than the overall WER for tasks such as spoken document retrieval and automatic content extraction.
4.4 Comparison with Related Methods

4.4.1 Maximum Entropy vs. Linear Interpolation

Compared to the back-off trigram model, which has about 500K parameters, the topic-conditional ME models introduce only about 15K additional parameters that modify probabilities of a few hundred words in the context of each topic. An alternative to this modeling approach is to partition the training data, build separate N-gram models $p_t$ for each topic $t$ and, since each topic N-gram is trained on a much smaller data set, interpolate this topic-specific model with a topic-independent model trained on all the data to obtain a smooth topic-dependent model. This is comparable to the approach described, e.g., in Clarkson & Robinson (1997), Iyer & Ostendorf (1996), Florian & Yarowsky (1999) and Martin, Liermann & Ney (1997).

We construct back-off unigram, bigram and trigram models specific to each topic using the partitioning of the 2.1 million word corpus used for the ME models as described in Section 4.2.3. We interpolate each topic-specific N-gram $p_t$ with the baseline (topic-independent) trigram model denoted as $p_3$ to obtain smooth topic-dependent N-gram models $p$ where

$$p(w_i|w_{i-2}, w_{i-1}) = \lambda p_3(w_i|w_{i-2}, w_{i-1}) + (1 - \lambda)p_t(w_i|w_{i-k}, w_{i-k+1}), \ 0 \leq \lambda \leq 1$$

for $k = 0, 1$ or 2 (for unigram, bigram and trigram models, respectively).

Usually, one would tune the interpolation coefficient $\lambda$ on some held-out set. In this case, however, we (cheat and) choose the interpolation weight to minimize the perplexity of the test set under each interpolated model. Table 4.10 shows the recognition performance of the interpolated models. The topic for each test utterance for the interpolated model is the same as the one used for the ME topic model.

Column 2 in Table 4.10 shows the number of parameters in different models. Even though the ME model has far fewer parameters than interpolation models, it outperforms the latter ones. It may thus be argued that the ME approach permits us to combine via unigram constraints as much effective information as one would get by interpolating topic-specific trigram models. This, we argue, is due to the systematic integration of topic-dependent and topic-independent constraints in our model.
<table>
<thead>
<tr>
<th>Model (N-gram cutoff)</th>
<th>#Params</th>
<th>Perplexity</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-Off (B≥1, T≥2)</td>
<td>499K</td>
<td>78.8</td>
<td>38.5%</td>
</tr>
<tr>
<td>Back-Off + Topic 1-gram</td>
<td>+70×11K</td>
<td>78.4</td>
<td>38.5%</td>
</tr>
<tr>
<td>Back-Off + Topic 2-gram</td>
<td>+70×26K</td>
<td>77.3</td>
<td>38.3%</td>
</tr>
<tr>
<td>Back-Off + Topic 3-gram</td>
<td>+70×55K</td>
<td>76.1</td>
<td>38.1%</td>
</tr>
<tr>
<td>ME-Topic (B≥1, T≥2)</td>
<td>+15K</td>
<td>73.5</td>
<td>37.8%</td>
</tr>
</tbody>
</table>

Table 4.10: Comparison with 70 interpolated topic N-gram models.

4.4.2 Topic Dependent Model vs. Cache-Based Model

The cache-based method (Kuhn & De Mori, 1990) is another widely used topic/domain adaptation technique for language modeling. We have implemented a cache-based language model for comparison with the topic-dependent maximum entropy model. Specifically, at each position \( i \) in the test corpus, we estimate a unigram probability \( p_c \) based on all the words seen so far in that particular conversation side. This unigram probability is then interpolated with the back-off trigram probability \( p_3 \) as

\[
p(w_i|w_{i-2}, w_{i-1}) = \lambda p_3(w_i|w_{i-2}, w_{i-1}) + (1 - \lambda)p_c(w_i), \quad 0 \leq \lambda \leq 1
\]

and used in rescoring.

Since the correct (spoken) words are not available to the recognizer, all the words in the N-best hypotheses for the preceding utterances in a conversation side are considered in estimating the cache model. This has the beneficial effect that words appearing in multiple hypotheses are likely to be correct and hence get counted multiple times, while words that do not appear in many hypotheses are potentially due to recognition errors and are naturally discounted.

Usually, one would tune this value of \( N \) as well as the interpolation coefficient on some held-out set. However, in this case as well, we choose the value of \( N \) (100) and the interpolation weight (0.1) of the cache language model to minimize the perplexity of the test set. The results in Table 4.11 show that although interpolation with the cache model results in reduced perplexity, there is no reduction in WER from cache model use. Cache models have, however, been shown to reduce WER on other tasks.
Table 4.11: Perplexity and WER of a cache-based model and a maximum entropy model with topic constraints.

While it is a little surprising to see the WER increase slightly\(^{11}\) with a cache model, this is perhaps explained by the relatively high error rates in the hypotheses from which the model is estimated. To support this argument, we compare the recognition output of the baseline trigram model and the cache-interpolated model of Table 4.11 in the following manner. Each recognition error is first marked as an insertion, substitution or deletion in the usual way. Next, each error token is marked as a repeated error if the word in question has undergone the same kind of error in the portion of the history used by the cache model. \(e.g.,\) if a word \(w\) in the reference transcription was deleted from the hypothesis for the first time in a conversation side at some position \(i\) and deleted again at position \(j > i\), then \(j\) is a case of a repeated deletion of \(w\), although \(i\) is not. Repeated insertions and substitutions of words are labeled in a similar manner. The word error rate is split into repeated error rate and non-repeated error rate (Table 4.12).

Table 4.12: Repeated errors vs. non-repeated errors.

Of the total error rate of 38.5\% for the trigram model, 35.8\% is accounted for by such repeated errors. For the cache-interpolated model, 36.3\% of the total error rate of 38.9\% is due to repeated errors. We find that compared to the baseline trigram model,

\(^{11}\) But only marginally significantly, with a P-value of 0.034.
the cache-based model has about a 0.6% higher rate of repeated errors. Comparing this to the overall increase in error rate of only 0.4%, it seems reasonable to conclude that a small reduction of non-repeated errors is offset by an increase in the number of repeated errors, pointing to the detrimental effect of using a cache model at high error rates.

4.5 Broadcast News Experiments

With the help of hierarchical training and divide-and-conquer, we also construct a topic-dependent ME model for a large task - the Broadcast News task. The overall experimental results have been presented in Wu & Khudanpur (2002). We provide more detailed results in this section.

We use the whole training set of 130 million words contained in about 125,000 stories. The vocabulary of 64K words is chosen to cover all words in the acoustic training data of 70 hours and those words that occur 11 times or more in the language model training text. All out-of-vocabulary words that count 0.33% of total tokens in the training data are mapped to a special token \textit{<unk>}. The acoustic model is a state-clustered cross-word triphone model with 39 dimension MFCC features. No speaker adaptation technique is used. The test set is the HUB-4 1996 eval set, containing about 570 utterances amounting to 22K words. In the experiments, word lattices for the utterances are first generated by the AT&T finite state machine (FSM) decoder (Mohri, Pereira & Riley, 2002), and then 100-best lists are created from lattices. The acoustic model scores and the language model scores of each of these 100-best hypotheses are obtained by aligning these hypotheses with the acoustic data using HTK. Finally, the 100-best lists are rescored using the new language models.

4.5.1 Performance of ME models

The baseline language model is a back-off trigram model with the bigram cut-off of 2 and the trigram-cut-off of 3. This is the trigram model with the lowest cut-offs that we can implement. The model contains about 64K unigrams, 3.5 million
bigrams and 5.5 million trigrams. The perplexity of the baseline model is 174.3 and the corresponding word error rate is 34.6% \(^{12}\) (Table 4.13). When computing the perplexity, 1.2% out-of-vocabulary words (OOVs) in the test data are mapped to the special token (\(<unk>\)). The topic-independent maximum entropy language model (with the same constraints as in the baseline model) is also built for comparison with the corresponding back-off model. As we have expected, this maximum entropy model almost duplicates the performance of the baseline back-off model (174.3 vs. 174.8 in perplexity and 34.6% vs. 34.5% in WER). The improvement from the topic-dependent model to the language model should be considered the result of adding topic dependencies.

To train the topic-dependent model, we obtain initial topic labels for documents in the training data from Florian & Yarowsky (1999), and then re-cluster the corpus into 100 topics by the K-means method mentioned in Section 4.2. In topic detection for the test utterances, we guess the best topic assignment using the same method as we did for the Switchboard experiments - assigning a fresh topic to each individual test utterance according the 10-best hypotheses of the current utterance and the four preceding ones (cf. Section 4.3.2). The topic-dependent model achieves a perplexity reduction of about 10% and a significant\(^ {13}\) WER reduction of about 0.6% absolute compared to the baseline trigram model. These results match those in the Switchboard experiments (a reduction of 7% for perplexity and a WER reduction of 0.7%) and show that the topic-dependent model is effective in different tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>Perplexity</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-off 3-gram ((B\geq2, \ T\geq3))</td>
<td>174.3</td>
<td>34.6%</td>
</tr>
<tr>
<td>ME 3-gram ((B\geq2, \ T\geq3))</td>
<td>174.8</td>
<td>34.5%</td>
</tr>
<tr>
<td>ME Topic 3-gram ((B\geq2, \ T\geq3))</td>
<td>156.8</td>
<td>34.0%</td>
</tr>
</tbody>
</table>

Table 4.13: Perplexity and WER of language models for the BN corpus

\(^{12}\) Although the perplexity of BN is higher than that of the Switchboard, its word error rate is relatively lower since the quality of speech is better than that of Switchboard.

\(^{13}\) With a p-value of \(10^{-4}\).
4.5.2 Comparison with Interpolated Models

We also build interpolated topic-dependent models for comparison as we have done for Switchboard. We interpolate topic-dependent unigram, bigram and trigram models with the baseline topic-independent trigram model. Topic tags for the test utterances are the same as those for the ME model. The results are presented in Table 4.14. The first 4 rows show the results of the baseline back-off model and the interpolated models. The last 2 rows duplicate the results of ME models from Table 4.13 for comparison. It is apparent that the performance of the best interpolation models (topic specific trigram models interpolating with the baseline model) almost reach the performance of the ME model. However, the size of interpolation models in a number of parameters increases about eightfold compared to the baseline trigram model, whereas that of the ME model increases by only 3%. Therefore, we argue that the ME method is more practical than interpolation for very large tasks when the size of the model is huge.

<table>
<thead>
<tr>
<th>Model (N-gram cutoff)</th>
<th>#Params</th>
<th>Ppl</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-Off 3-gram</td>
<td>9.1M</td>
<td>174.3</td>
<td>34.6%</td>
</tr>
<tr>
<td>BO + Topic 1-gram</td>
<td>+100×60K</td>
<td>167.6</td>
<td>34.5%</td>
</tr>
<tr>
<td>BO + Topic 2-gram</td>
<td>+100×400K</td>
<td>162.3</td>
<td>34.3%</td>
</tr>
<tr>
<td>BO + Topic 3-gram</td>
<td>+100×600K</td>
<td>156.1</td>
<td>34.1%</td>
</tr>
<tr>
<td>ME 3-gram</td>
<td>9.1M</td>
<td>174.8</td>
<td>34.5%</td>
</tr>
<tr>
<td>ME Topic</td>
<td>+250K</td>
<td>156.8</td>
<td>34.0%</td>
</tr>
</tbody>
</table>

Table 4.14: Comparison with 100 interpolated topic N-gram models.

4.6 Summary

We have constructed topic-dependent models using the maximum entropy method for a small task of telephone conversations (Switchboard) and a large task of Broadcast News. By incorporating a small number of topic constraints with N-grams, we have successfully reduced both the perplexity (7%) for the Switchboard and 10% for
the Broadcast News) and WER (0.7% absolute for the Switchboard and 0.6% absolute for the Broadcast News). We have studied the influence of different ways of topic assignment in recognition. The experimental results show that using N-best hypotheses causes little degradation (in perplexity and WER) and assigning topics at utterance level is slightly better than that at the conversation level. We have also compared the maximum entropy method with linear interpolation. The maximum entropy method is more efficient than linear interpolation in combining topic dependencies with N-grams. Finally, we have compared the topic-dependent models with cache-based ones and found that the former is better than the latter in reducing WER when the baseline is poor.