ABSTRACT

Including phrases in the vocabulary list can improve n-gram language models used in speech recognition. In this paper, we report results of automatic extraction of phrases from the training text using frequency, likelihood, and correlation criteria. We show how a language model built from a vocabulary that includes useful phrases can systematically improve language model perplexity in a natural language call-routing task and the 20K-Nov92 Wall Street Journal evaluation. We also discuss the impact of such phrase-based language models on recognition word error rate.

Keywords: language models, phrase models, automatic phrase extraction.

1. INTRODUCTION

Words are commonly used as the basic lexical units in standard language models for automatic speech recognition (ASR). However, the inclusion of recurrent phrases into the vocabulary list has been proposed by various researchers to reduce the perplexity of the language model and to improve ASR performance [1,2,3,4].

One way that phrases improve the language model is by capturing a longer context length, similar to the effects of using n-grams of variable length [5]. Certain phrases may have meanings which are different from those of the individual words, i.e. they are collocations. These phrases (e.g. “write_offs”) may have different statistical properties from the component words (“write,” “off”). Phrase-based language models may also improve ASR accuracy by allowing for better acoustic modeling of inter-word boundaries (e.g. “in-the” or “you-all”) and the utilization of inter-word pronunciation variants.

Although phrases may be manually selected and added to the lexicon, it is of interest to be able to automatically extract phrases from training text data based on certain measures such as the change in language model perplexity after the merging of words into a phrase. We perform a systematic study of automatically selecting phrases from training text using various criteria, and discuss the effects on the language model perplexity with respect to held-out data and the word recognition error rate. Two corpora of different sizes are studied. One task involves a natural language call-routing task, which has a 2K vocabulary and 200K-word training text [6]. The other task is the Wall Street Journal (WSJ) corpus with a 20K vocabulary and 37M-word training text, of which 10% of the sentences are held out for calculating test set perplexity. We address the issue of where phrase modeling is most likely to make an impact on speech recognition.

2. PHRASE SELECTION

2.1 Phrase Selection Algorithm

Phrases are automatically extracted from the training text. The goal is to select a set of phrases which when added to the lexicon reduces the perplexity of the language model. Unfortunately, finding an optimal set of phrases is an NP-complete problem, so we resort to a greedy algorithm that selects the best candidate iteratively. In each iteration, word pairs such as “new_york” are merged into phrases (e.g. “new_york”) and added to the lexicon. In addition, we select more than one candidate in each iteration to reduce the number of iterations required to obtain a certain number of phrases. The candidate phrases have to meet certain requirements:

1) The coverage of the new phrase “u_v” is large enough, i.e. the number of occurrence of the phrase has to exceed a count threshold: \( N(u_v) > C_{th} \). For the call-routing task, we use a threshold \( C_{th}=20-50 \), and for the WSJ task, \( C_{th}=250 \).

2) In each iteration only the best \( n_b \) phrase candidates based on a particular selection criterion are considered for addition to the lexicon. They are ranked according to the phrase selection criterion.

3) For the log likelihood criteria (mentioned below), the change in overall log likelihood for the training text has to be positive.

4) Phrases which may conflict with higher ranking candidates are removed, e.g. if “u_v” has been chosen to be merged, then all phrase candidates of lower rank ending in “u” or beginning in “v” are removed in the current iteration.

Phrases meeting all these requirements are added to the lexicon and substituted into the corpus. Then a new iteration is entered. Note that a previously added phrase (e.g. “new_york”) is considered a lexical entry just like any other word and can be further merged with other
lexical entries to form a longer phrase (e.g., “new_york_stock”). If the phrase count of “new_york” falls below a threshold because a fraction was merged with other words, the phrase is split into its component words “new” and “york.”

2.2 Phrase Selection Criteria

Given a training text with a total of number of words \(N_T\), word counts \(N(w)\) and word-pair counts \(N(w_1, w_2)\), we used different measures to evaluate a particular phrase candidate “\(u\) \(v\)” for addition to the lexicon:

1) Frequency of word pair: \(N(u,v)/N_T\)

2) Change in unigram log likelihood \((\Delta F_{uni})\) due to addition of phrase into lexicon, where the unigram log likelihood of the training text is given by:

\[
F_{uni} = \sum_w N(w) \log \frac{N(w)}{N_T}
\]

3) Change in bigram log likelihood \((\Delta F_{bi})\), where the bigram log likelihood is given by:

\[
F_{bi} = \sum_{w_1, w_2} N(w_1, w_2) \log \frac{N(w_1, w_2)}{N(w_1)}
\]

It can be shown that the change in unigram log likelihood is related to the mutual information as follows:

\[
\Delta F_{uni} = I_M(u, v) + [N(u) - N(u, v)] \log \left(1 - \frac{N(u, v)}{N(u)}\right)
\]

\[
+ [N(v) - N(u, v)] \log \left(1 - \frac{N(u, v)}{N(v)}\right)
\]

where

\[
I_M(u, v) = N(u,v) \log \left(\frac{N(u,v)N_T}{N(u)N(v)}\right)
\]

If \(N(u) \gg N(u,v)\) and \(N(v) \gg N(u,v)\) or if \(N(u) \approx N(u,v)\) and \(N(v) \approx N(u,v)\), the change in unigram likelihood will be approximately equal to the mutual information. Thus word pairs which increases the unigram likelihood have high mutual information.

In some experiments we also used a type of correlation coefficient [3] for comparison:

\[
\rho_{u,v} = \frac{P(u,v)}{P(u) + P(v)},
\]

where \(P(u)\) and \(P(v)\) are the respective probabilities of the words “\(u\)” and “\(v\).” This correlation measure is similar to the mutual information but the denominator term is a sum, rather than product, of the two probabilities. This measure has a maximum value of 0.5, when \(P(u,v) = P(u) = P(v)\), i.e. when words “\(u\)” and “\(v\)” always occur together.

Note that to calculate the change in unigram log likelihood \((\Delta F_{uni})\), bigram counts are needed; for \(\Delta F_{bi}\), trigram counts are needed. For change in trigram log likelihood, 4-gram counts would have been needed.

3. RESULTS

3.1 Natural Language Call-routing Task

Results from the natural-language call-routing task show monotonically decreasing bigram perplexity as a function of the number of phrases, with a reduction in perplexity of about 12% but a slight increase in trigram perplexity of about 2%. Note that the perplexity calculation was normalized to the original number of words.

Figure 1 shows the reduction in bigram perplexity as a function of the number of new phrases. All three phrase selection criteria gave similar results, with the bigram

![Figure 1: Bigram perplexity versus the number of phrases added to the lexicon for call-routing task.](image1)

![Figure 2: Word error rate as a function of the number of phrases added to the lexicon for call-routing task.](image2)
likelihood measure yielding slightly better results. There is a rapid decrease in perplexity as new phrases are initially added to the lexicon, but the effect levels off as more phrases are added. Examples of phrases extracted by the algorithm include: “I’d_like_to,” “I_have_a_question,” and “checking_accounts.”

The recognition word error rate for this telephone-based application is reduced by 30% from about 37% to 26% as shown in Figure 2. The unigram likelihood criterion is slightly better than the bigram likelihood criterion. For this reason and because bigram likelihood calculation is computationally expensive for a large training text, we chose not to use the bigram likelihood criterion for the Wall Street Journal task.

3.2 Wall Street Journal Task

The results for the larger WSJ corpus show a reduction in bigram perplexity of about 9% and trigram perplexity of about 5%. Figure 3 shows the decrease in the trigram perplexity as a function of the number of phrases. There is a rapid perplexity decrease followed by a leveling off as the number of phrases increases. The word-pair frequency criterion for selecting phrases gives the best results, followed by unigram likelihood and correlation criteria.

In speech recognition experiments, only a modest reduction in word error rate (WER) from 9.3% to 9.2% for the 20K-Nov92 WSJ evaluation was achieved. Addition of optional silences in the pronunciation lexicon resulted in a further reduction in the WER to 9.1%.

3.3 Phrase characteristics

The phrases selected using different phrase selection criteria have distinct characteristics. Phrases selected using word-pair frequency include phrases like “of_the” which occur frequently simply because they consist of function words like “of” and “the” which by themselves have high probabilities. The component words in the phrases selected by the word-pair frequency criterion need not have greater affinity for one another than for other words.

In contrast, the correlation criterion is insensitive to the number of occurrences of the phrase. Instead, phrases selected have component words which are correlated to one another. For example, the word “los” is almost always followed by “angeles” and the word “angeles” is almost always preceded by “los” in the training text. The correlation criterion will select such phrases even if the phrases have low frequency.

<table>
<thead>
<tr>
<th>Word pair Frequency</th>
<th>Unigram Likelihood</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>of_the</td>
<td>u.s.</td>
<td>los_angeles</td>
</tr>
<tr>
<td>in_the</td>
<td>million_dollars</td>
<td>hong_kong</td>
</tr>
<tr>
<td>million_dollars</td>
<td>nineteen_eighty</td>
<td>las_vegas</td>
</tr>
<tr>
<td>u.s.</td>
<td>new_york</td>
<td>palo_alto</td>
</tr>
<tr>
<td>nineteen_eighty</td>
<td>billion_dollars</td>
<td>donaldson_lufkin</td>
</tr>
<tr>
<td>for_the</td>
<td>one_hundred</td>
<td>kohberg_kravis</td>
</tr>
<tr>
<td>to_the</td>
<td>more than</td>
<td>coca_cola</td>
</tr>
<tr>
<td>on_the</td>
<td>vice_president</td>
<td>hewlett_packard</td>
</tr>
<tr>
<td>one_hundred</td>
<td>a_share</td>
<td>burnham_lambard</td>
</tr>
<tr>
<td>that_the</td>
<td>has_been</td>
<td>et_cetera</td>
</tr>
</tbody>
</table>

Both correlation and frequency are important with the unigram likelihood criterion. It is different from the word-pair frequency in preferring phrases such as “new_york” and “vice_president” whose component words have higher correlation than “of_the” but do not occur as frequently. On the other hand, it also prefers phrases which have high counts, so it does not highly rank phrases like “kohberg_kravis” which occur only about 600 times in the training text of 37M words.

Phrases selected by the correlation criterion may seem to be more natural phrase choices. However, the actual perplexity results reported earlier indicates that the correlation criterion is not as good as the unigram likelihood or the word frequency criteria in reducing the perplexity. Although correlated words found through the correlation criterion are natural candidates for merging into phrases, in a large vocabulary system, these phrases may have low frequency and thus may not affect the perplexity of the language model significantly.
Figure 4 shows the coverage of the new phrases in the test set used for speech recognition experiment. Coverage is defined to be the percentage of the new phrases which actually appear in the test set at least once. As Figure 4 shows, the first few hundred phrases selected using the word-pair frequency and unigram likelihood criteria are very likely to be used. As an example, more than 50% of the first 200 phrases appear at least once in the test set. As more phrases are added to the lexicon, the fraction of them which actually appear in the test set decreases.

In contrast, phrases selected using the correlation criterion tend to have a fairly uniform but low coverage. This is reasonable since no frequency information was considered in choosing these phrases, so they represent a random sample of words. Only about 20% of the phrases selected by the correlation criterion appear in the test set.

4. DISCUSSION

Results from using phrase-based language models in the call-routing task and the WSJ task reveal interesting insights. First of all, adding phrases reduces the language model bigram perplexities and the word recognition error rates in both tasks. The reduction in perplexity tends to be more rapid at the beginning and level off as more phrases are added to the lexicon. However, the recognition word error rate reduction was significantly larger in the call-routing task (from 37% to 26%) than in the WSJ task (9.3% to 9.1%).

There are several possible reasons that phrase models have greater impact on one task than the other. The tasks are different in several aspects. The WSJ has a large vocabulary with a high perplexity (trigram perplexity ~ 130) whereas the call-routing task has a much smaller vocabulary and perplexity (trigram perplexity ~ 30). The amount of training data for the WSJ is much larger. The conditions for the two tasks are also different. In the WSJ task, the recordings are made of clean, read speech recorded with microphones. In the call-routing task, the speech is spontaneous and is recorded over the telephone.

Phrase models appear to make a greater impact on improving the recognition word error rate in an acoustically challenging situation where there has not been a large amount of training data for the language model and where the perplexity is low. The utterances for the call-routing task tend to have stereotypical constructs with recurrent phrases such as “I want to,” “Can you please get me,” “credit card services,” “auto loan department,” etc. These stereotypical constructs are very naturally modeled as phrases in the language model. This style is very different from the written style of the WSJ where sentences have been carefully edited to maintain variety and richness in vocabulary and construction. In this case, the most frequent phrases are those of function words such as “of the” or “in the.” Longer phrases such as “new york stock exchange” can also be extracted but do not necessarily have a high probability of appearing in the test set.

5. ACKNOWLEDGEMENTS

We acknowledge Brett Diamond for his work on the USAA task. We thank Stefan Ortmanns for helpful discussions.

6. REFERENCES