On Using Written Language Training Data for Spoken Language Modeling

R. Schwartz, L. Nguyen, F. Kubala, G. Chou, G. Zavaliagkos†*, J. Makhoul*

BBN Systems and Technologies
Cambridge, MA 02138
†Northeastern University

ABSTRACT

We attemped to improve recognition accuracy by reducing the inadequacies of the lexicon and language model. Specifically we address the following three problems: (1) the best size for the lexicon, (2) conditioning written text for spoken language recognition, and (3) using additional training outside the text distribution. We found that increasing the lexicon 20,000 words to 40,000 words reduced the percentage of words outside the vocabulary from over 2% to just 0.2%, thereby decreasing the error rate substantially. The error rate on words already in the vocabulary did not increase substantially. We modified the language model training text by applying rules to simulate the differences between the training text and what people actually said. Finally, we found that using another three years' of training text - even without the appropriate preprocessing, substantially improved the language model. We also tested these approaches on spontaneous news dictation and found similar improvements.

1. INTRODUCTION

Speech recognition accuracy is affected as much by the language model as by the acoustic model. In general, the word error rate is roughly proportional to the square root of the perplexity of the language model. In addition, in a natural unlimited vocabulary task, a substantial portion of the word errors come from words that are not even in the recognition vocabulary. These out-of-vocabulary (OOV) words have no chance of being recognized correctly. Thus, our goal is to estimate a good language model from the available training text, and to determine a vocabulary that is likely to cover the test vocabulary.

The straightforward solution to improving the language model might be to increase the complexity of the model (e.g., use a higher order Markov chain) and/or obtain more language model training text. But this by itself will not necessarily provide a better model, especially if the text is not an ideal model of what people will actually say. The simple solution to increase the coverage of the vocabulary is to increase the vocabulary size. But this also increases the word error rate and the computation and size of the recognition process.

In this paper we consider several simple techniques for improving the power of the language model. First, in Section 3, we explore the effect of increasing the vocabulary size on recognition accuracy in an unlimited vocabulary task. Second, in Section 4, we consider ways to model the differences between the language model training text and the way people actually speak. And third, in Section 5, we show that simply increasing the amount of language model training helps significantly.

2. THE WSJ CORPUS

The November 1993 ARPA Continuous Speech Recognition (CSR) evaluations was based on speech and language taken from the Wall Street Journal (WSJ). The standard language model training text was estimated from about 35 million words of text extracted from the WSJ from 1987 to 1989. The text was normalized (preprocessed) with a model for what words people use to read open text. For example, "$234.56" was *always* assumed to be read as "two hundred thirty four dollars and fifty six cents". "March 13" was always normalized as "March thirteenth" – not "March the thirteenth", nor "March thirteen". And so on.

The original processed text contains about 160,000 unique words. However, many of these are due to misspellings. Therefore, the test corpus was limited to those sentences that consisted only of the most likely 64,000 words. While this vocabulary is still quite large, it has two beneficial effects. First, it greatly reduces the number of misspellings in the texts. Second, it allows implementations to use 2-byte data fields to represent the words rather than having to use 4 bytes.

The "standard" recognition vocabulary was defined as the most likely 20,000 words in the corpus. Then, the standard language model was defined as a trigram language model estimated specifically for these 20K words. This standard model, provided by Lincoln Laboratory, was to be used for the controlled portion of the recognition tests. In addition, participants were encouraged to generate an improved language model by any means (other than examining the test data).

3. RECOGNITION LEXICON

We find that, typically, over 2% of the word occurrences in a development set are not included in the standard 20K-word vocabulary. Naturally, words that are not in the vocabulary cannot be recognized accurately. (At best, we might try to detect that there is one or more unknown words at this point in a sentence, and then attempt to recognize the phoneme sequence, and then guess a possible letter sequence for this phoneme sequence. Unfortunately, in English, even if we could recognize the phonemes perfectly, there are many valid ways to spell a particular phoneme sequence.) However, in addition to this word not being recognized, we often see that one or two words adjacent to this missing word are also misrecognized. This is because the recognition, in choosing a word in its vocabulary, also now has the wrong context for the following or preceding words. In general, we find that the word error rate increases by about 1.5 to 2 times the number of out-of-vocabulary (OOV) words.

One simple way to decrease the percentage of OOV words is to increase the vocabulary size. But which words should be added? The obvious solution is to add words in order of their relative frequency within the full text corpus. There are several problems that might result from this:

1. The vocabulary might have to be extremely large before the OOV rate is reduced significantly.
2. If the word error rate for the vast majority of the words that are already in the smaller vocabulary increased by even a small amount, it might offset any gain obtained from reducing the OOV rate.
3. The language model probabilities for these additional words would be quite low, which might prevent them from being recognized anyway.

We did not have phonetic pronunciations for all of the 64K words. We sent a list of the (approximately 34K) words for which we had no pronunciations to Boston University. They found pronunciations for about half (18K) of the words in their (expanded Moby) dictionary. When we added these words to our WSJ dictionary, we had a total of 50K words that we could use for recognition.

The following table shows the percentage of OOV words as a function of the vocabulary size. The measurement was done on the WSJ1 Hub1 "20K" development test which has 2,464 unique words with the total count of 8,227 words. Due to the unavailability of phonetic pronunciations (mentioned above), the final vocabulary size would be the second column.

Top N	Vocab.	#OOV	%
20k	19998	187	2.27
30k	28247	85	1.03
40k	35298	39	0.47
48k	40213	14	0.17
50k	41363	12	0.15
64k	48386	1	0.01

We were somewhat surprised to see that the percentage of OOV words was reduced to only 0.17% when the lexicon included the most likely 40K words – especially given that many of the most likely words were not available because we did not have phonetic pronunciations for them. Thus, it was not necessary to increase the vocabulary above 40K words.

The second worry was that increasing the vocabulary by too much might increase the word error rate due to the increased number of choices. For example, normally, if we double the vocabulary, we might expect an increase in word error rate of about 40%! So we performed an experiment in which we used the standard 20K language model for the 5K development data. We found, to our surprise, that the error rate increased only slightly, from 8.7% to 9.3%. Therefore, we felt confident that we could increase the vocabulary as needed.

We considered possible explanations for the small increase in error due to a larger vocabulary. We realized that the answer was in the language model. In the first case, when we just increase the vocabulary, the new words also have the same probability in the language model as the old words. However, in this case, all the new words that were added had lower probabilities (at least for the unigram model) than the existing words. Let us consider two possibilities that we would not falsely substitute a new word for an old one. If the new word were acoustically similar to one of the words in the test (and therefore similar to a word in the original vocabulary, then the word would be correctly recognized because the original word would always have a higher language model probability. If, on the other hand, the new word were acoustically very different from the word being spoken, then we might expect that our acoustic models would prevent the new word from being chosen over the old word. While the argument makes some sense, we did not expect the loss for increasing the vocabulary from 5K words to 20K words to be so small.

Finally, the third question is whether the new words would be recognized when they did occur, since (as mentioned above) their language model probabilities were generally low. In fact, we found that, even though the error rate for these new words was higher than for the more likely words, we were still able to recognize about 50% to 70% of them correctly, presumably based largely on the acoustic model. Thus, the net effect of this was to reduce the word error rate by about 1% to 1.5%, absolute.

4. MODELING SPOKEN LANGUAGE

Another effect that we worked on was the difference between the processed text, as defined by the preprocessor, and the words that people actually used when reading WSJ text. In the pilot WSJ corpus, the subjects were prompted with texts that had already been "normalized", so that there was no ambiguity about how to read a sentence. However, in the WSJ1 corpus, subjects were instructed to read the original texts and to say whatever seemed most appropriate to them. Since the WSJ1 prompting texts were not normalized to deterministic word sequences, subjects showed considerable variability in their reading of the prompting text.

However, the standard language model was derived from the normalized text produced by the preprocessor. This resulted in a mismatch between the language model and the actual word sequences that were spoken. While the preprocessor was quite good at predicting what people said most of the time, there were several cases where people used different words than predicted. For example, the preprocessor predicted that strings like "$234" would be read as "two hundred thirty four dollars". But in fact, most people read this as "two hundred AND thirty four dollars". For another extreme example, the preprocessor's prediction of "10.4" was "ten point four", but the subject (in the WSJ1 development data) read this as "ten and four tenths". There were many other similar examples.

The standard model for the tests was the "nonverbalized punctuation" (NVP) model, which assumes that the readers never speak any of the punctuation words. The other model that had been defined was the "verbalized punctuation" (VP) model, which assumed that *all* of the punctuation was read out loud. This year, the subjects were instructed that they were free to read the punctuation out loud or not, in whatever way they feel most comfortable. It turns out that people didn't verbalize most punctuation. However, they regularly verbalized quotation marks in many different ways that were all different than the ways predicted by the standard preprocessor.

There were also several words that were read differently by subjects. For example, subjects pronounced abbreviations like, "CORP." and "INC.". While the preprocessor assumed that all abbreviations would be read as full words.

We used two methods to model the ways people actually read text. The simpler approach was to include the text of the acoustic training data in the language model training. That is, we simply added the 37K sentence transcriptions from the acoustic training to the 2M sentences of training text. The advantage of this method is that it modeled what people actually said. The system was definitely more likely to recognize words or sequences that were previously impossible. The problem with this method was that the amount of transcribed speech was quite small (about 50 times smaller) compared to the original training text. We tried repeating the transcriptions several times, but we found that the effect was not as strong as we would like.

A more powerful approach was to simulate the effects of the different word choices by simple rules which were applied to all of the 35M words of language training text. We chose to use the following rules:

Preprocessed Text	Simulated Text
HUNDRED **[number]**	HUNDRED AND **[number]**
ONE HUNDRED	A HUNDRED
ONE DOLLAR	A DOLLAR
ZERO POINT **[number]**	POINT **[number]**
AND ONE HALF	AND A HALF
AND ONE QUARTER	AND A QUARTER

Thus, for example, if the sentence consists of the pattern "hundred twenty", we repeated the same sentence with "hundred AND twenty".

The result was that about one fifth of the sentences in the original corpus had some change reflecting a difference in the way subjects read the original text. Thus, this was equivalent in weight to an equal amount of training text to the original text.

We found that this preprocessing of the text was sufficient to cover most of those cases where the readers said things differently than the predictions. The recognition results showed that the system now usually recognized the new word sequences and abbreviations correctly.

5. INCREASING THE LANGUAGE MODEL TRAINING

While 35M words may seem like a lot of data, it is not enough to cover all of the trigrams that are likely to occur in the testing data. So we considered other sources for additional language modeling text. The only easily accessible data available was an additional 3 years (from 1990-1992) of WSJ data from the TIPSTER corpus produced by the Linguistic Data Consortium (LDC).

However, there were two problems with using this data. First, since the test data was known to come from 1987-1989, we were concerned that this might actually hurt performance due to some differences in the topics during that 3-year period. Second, this text had not been normalized with the preprocessor and we did not have available to us the preprocessor that was used to transform the raw text into word sequences.

We decided to use the new text with minimal processing. The text was filtered to remove all tables, captions, numbers, etc. We replaced each initial example of double-quote (") with "QUOTE and the matching token with "UNQUOTE or "ENDQUOTE, which were the most common ways these words were said. No other changes were made. We just used the raw text as it was. One benefit of this was that abbreviations were left as they appeared in the text rather than expanded. Any numbers, dates, dollar amounts, etc, were just considered "unknown" words, and did not contribute to the training. We assumed that we had sufficient examples of numbers in the original text.

We found that adding this additional language training data reduced the error by about 7% of the error, indicating that the original 35 million words was not sufficient for the models we were using. Thus, the addition of plain text, even though it was from a different three years, and had many gaps due to apparent unknown words, still improved the recognition accuracy considerably.

6. RESULTS

The following table shows the benefit of the enlarged 40K lexicon and the enhanced language model training on the OOV rate and the word error for the development test and the evaluation test.

	% OOV		% Word Error	
Test Set	20K	40K	20K	40K
Development	2.27	0.17	16.4	12.9
Evaluation	1.83	0.23	14.2	12.2

Surprisingly, the addition of three year's LM training (from a period post-dating the test data) improved performance on the utterances that were completely inside the vocabulary. Evidently, even the common trigrams are poorly trained with only the 35 million word WSJ0 corpus. Overall, our modifications to the lexicon and grammar training reduced the word error by 14–22%.

7. Spontaneous Dictation

Another area we investigated was spontaneous dictation. The subjects were primarily former or practicing journalists with some experience at dictation. They were instructed to dictate general and financial news stories that would be appropriate for a newspaper like WSJ. In general, the journalists chose topics of recent interest. This meant that the original language model was often out of date for the subject. As a result, the percentage of OOV words increased (to about 4%), and the language model taken from WSJ text was less appropriate.

The OOV words in the spontaneous data were more likely to be proper nouns from recent events that were not covered by the LM training material. To counter this, we added all (1,028) of the new words that were found in the spontaneous portion of the acoustic training data in WSJ1. This mostly included topical names (e.g., Hillary Rodham, NAFTA, etc.).

In order to account for some of the differences between the read text and the spontaneous text, and to have language model probabilities for the new words, we added the training transcriptions of the spontaneous dictation (about 8K sentences) to the LM training as well.

New weights for the new language model, HMM, and Segmental Neural Network were all optimized on spontaneous development test data. The table below shows that the OOV remains near 1% even after the enlargement to a 41K lexicon.

	% OOV			% Word Error	
Test Set	20K	40K	41K	20K	41K
Development	2.9	1.4	0.8	–	21.7
Evaluation	4.8	1.9	1.5	24.7	19.1

As can be seen, increasing the vocabulary size from 20K to 40K significantly reduced the OOV rate. It is important to point out that in this case, we did not have the benefit of a word frequency list for spontaneous speech, and that the source of speech had an unlimited vocabulary. So the reduction in OOV rate is certainly a fair – if not pessimistic – estimate of the real benefit from increasing the vocabulary. Adding the few new words observed in the spontaneous speech also helped somewhat, but not nearly as much. The sample of only 8,000 sentences is clearly not sufficient to find all the new words that people might use. Presumably, if the sample of spontaneous speech were large enough to derive word frequencies, then we could choose a much better list of 40K words with a lower OOV rate.

Overall, the 41K trigram reduces the word error by 23% over the 20K standard trigram on the November '93 CSR S9 evaluation test. We estimate that more than half of this gain was due to the decreased percentage of OOV words, and the remainder was due to the increased language model training, including specific examples of spontaneous dictation.

8. CONCLUSIONS

We found the following interesting results:

- Expanding the vocabulary with less frequent words does not substantially increase the word error on those words already in the vocabulary, but does eliminate many errors due to OOV words.
- Doubling the amount of language model training text improves the language model, even though the text comes from different years than the test, and even though the text was not preprocessed into proper lexical forms.
- It is possible to improve the quality of the language modeling text by modeling the differences between the

predicted reading style and some examples of actual transcriptions.

- Increasing the vocabulary size and language training had a bigger effect on spontaneous speech than it did for read speech.

9. ACKKNOWLEDGEMENT

This work was supported by the Advanced Research Projects Agency and monitored by the Office of Naval Research under contract No. N00014-92-C-0035.

References

1. Bates, M., R. Bobrow, P. Fung, R. Ingria, F. Kubala, J. Makhoul, L. Nguyen, R. Schwartz, D. Stallard, "The BBN/HARC Spoken Language Understanding System", *Proc. of IEEE ICASSP-93*, Minneapolis, MN, April 1993, pp. 111-114, vol. II.

2. Placeway, P., R. Schwartz, P. Fung, L. Nguyen, "The Estimation of Powerful Language Models from Small and Large Corpora", *Proc. of IEEE ICASSP-93*, Minneapolis, MN, April 1993, vol. II, pp. 33-36.