Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data

Ronald Rosenfeld

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

I study the effect of various types and amounts of North American Business language data on the quality of the derived vocabulary, and use my findings to derive an improved ranking of the words, using only 19% of the NAB corpus. I then study the conflicting effects of increased vocabulary size on a speech recognizer's accuracy, and use the result to pick an optimal vocabulary size. A similar analysis of ngram coverage yields a very different outcome, with the best system being the one based on the most data.

1. Vocabulary Optimization

1.1. OOV curve minimization

Since Out-Of-Vocabulary (OOV) rate directly affects Word Error Rate, with every OOV word in the test data resulting in at least one (and often more) recognition errors, I set out to minimize the expected OOV rate of the test data. More generally, my goal was to understand how availability of various types and amounts of training data, from various time periods, affects the quality of the derived vocabulary [1]. Given a collection of training data, I sought to create an ordered word list with the lowest possible OOV curve, such that, for any desired vocabulary size V, a minimum-OOV-rate vocabulary could be derived by taking the first V words in that list. Viewed this way, the problem becomes one of estimating unigram probabilities of the test distribution, and then ordering the words by these estimates.

The test set consisted of 1.4M words worth of North American Business news. The training data was the 227M-word NAB corpus (see [Roseneld 95] for details). In all studies, except where otherwise noted, the word list was ordered by decreasing frequency in the appropriate subset of the training data.

I first set out to measure the effect on OOV rate of the *seasonality* of the training data, namely the time of year from which it is drawn. For each month of the year, I created a word list based on some 9MW of training data from that month. The test data was drawn from 4/94, so a seasonal effect might reduce the OOV rate of training data from this or adjacent months. As Figure 1 shows, no such effect was found.

Next I measured the correlation of OOV rate with the *amount* of training data. I added training data in increments of 5MW, and measured the impact on OOV rate. I added data in decreased order of recency, so as not to confound the effect of the amount of data with that of its recency. Figure 2 shows my findings. As expected, more training data results in lower OOV rates. But improvement slows down considerably after 30MW–50MW. Next, I studied the effect of

[1] The vocabulary thus derived is *static*. It can serve as the initial vocabulary, to be optionally extended at runtime based on the words encountered in the test data.

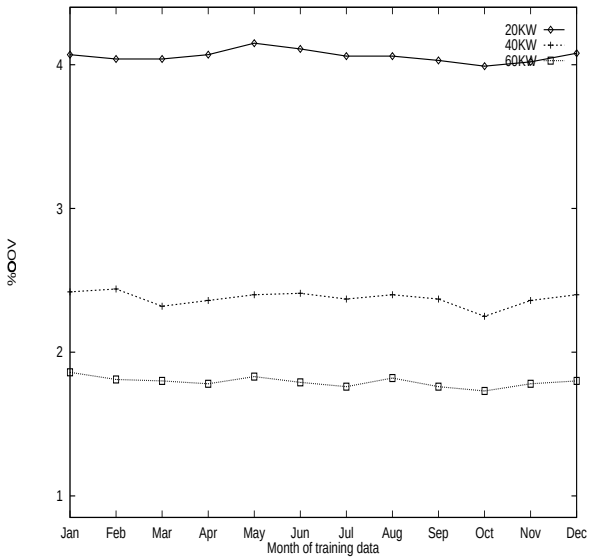

Figure 1: Month from which training data is drawn has no effect on OOV rates (test data is from April).

recency of the training data. Figure 3 shows OOV rates based on similar amounts of training data (about 5MW), but from different time periods. Time indeed makes a difference, albeit slowly. Over a period of 2 years, the 20KW (60KW) OOV rate degraded by 5% (15%). Over 4.5 years, it degraded by 11% (24%).

The difference that the *source* of the training data can make is evident in Figure 4. An OOV curve based on the Wall-Street-Journal(1990) part of the data (10MW), is lower than that based on the San-Jose-Mercury(1991) part (11MW), even though the latter is larger and more recent.

Next, I accumulated data starting from the most recent period and going backwards in time. Given the inherent tradeoff between the amount of data and its recency and source, I hypothesized a U-shape OOV curve, which was indeed achieved as can be seen in Figure 5 (the last datapoint is based on the entire 227MW NAB training corpus). The peak was achieved at about 40MW. It is interesting that the best overall coverage was obtained using only 19% of the available training data!

If recent data is more useful, can we benefit from emphasizing it? Several such attempts failed. The only one that was mildly successful was based on a "leaky capacitor" model of word probabilities. Discounting the word counts by 1% every week reduced the OOV

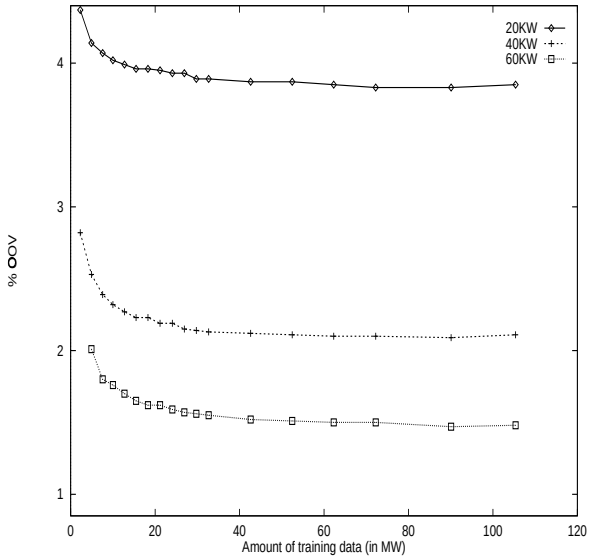

Figure 2: More training data results in lower OOV rates, but mostly up to 30MW–50MW

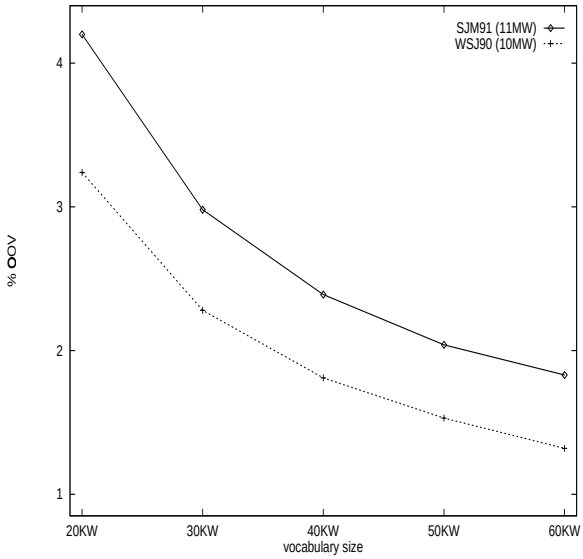

Figure 4: The source of the training data makes a big difference in OOV rates.

rate very slightly for vocabulary sizes in the range 20KW–50KW, but not at the 60KW level.

1.2. Vocabulary size optimization

Increasing the vocabulary of a speech recognition system has two conflicting effects. On one hand, it reduces the OOV rate, thereby helping to recover OOV related recognition errors. On the other hand, the added lexical entries increase the average acoustic confusability of words, resulting in new recognition errors.

To quantify these two effects, I ran two controlled experiments on the CSR 1994 acoustic development test set. In the first, I compared two systems that differed only in their lexicon. The first system had a lexicon of 58K words. The second was based on the top 20K words of the 58K lexicon, supplemented with all the test set words that were in the 58K lexicon. Thus the two lexicons had identical coverage of the test set, but very different overall sizes. The 58KW system resulted in 0.6 points higher WER. I interpreted the difference as resulting solely from the increased acoustic confusability. Assuming that acoustic confusability grows roughly linearly with vocabulary size, I arrived at a slope of +0.16 WER points per 10KW increase in the vocabulary. Alternatively, assuming that acoustic confusability grows logarithmically with vocabulary size, I arrived at a slope of +0.39 WER points per doubling of the vocabulary size.

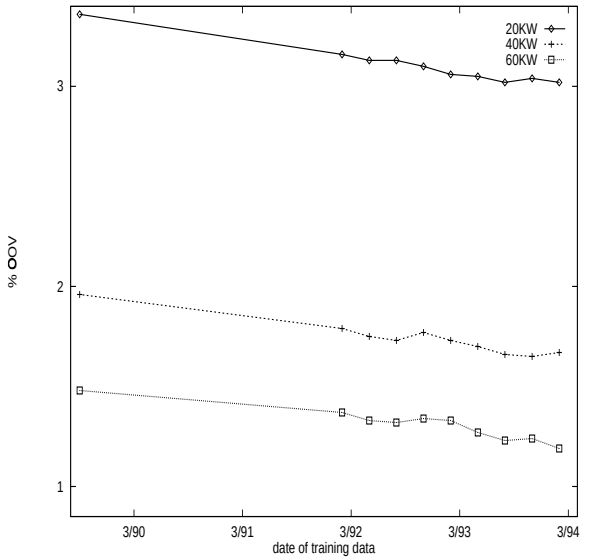

Figure 3: More recent training data results in lower OOV rates.

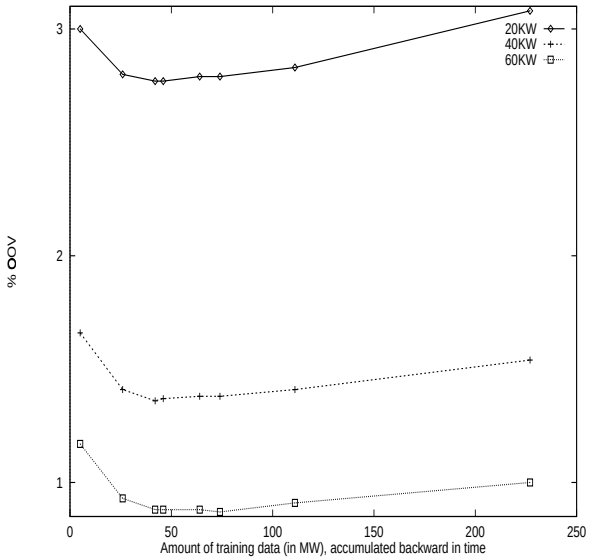

Figure 5: Best OOV rates are achieved with only 19% of the training data!

In the second experiment, I again compared two systems differing only in their lexicon. One system used the 58K lexicon; the other used the same 20K lexicon as above (unsupplemented), and had a 1.79% higher OOV rate. Thus, this time the lexicons differed in both size and test-set coverage. The 20KW system had a 1.55 points higher WER. Factoring in the 0.6 points WER reduction due to the reduced confusability, I corrected the effective difference to 2.15 WER points. Assuming that OOV-related errors are linear with the OOV rate, I arrived at a slope of –1.2 WER points per OOV-point, or an average of 1.2 word recognition errors per OOV word .

As we increase vocabulary size, OOV rate decreases at an ever slower rate. For any OOV curve, there is a point at which the savings due to reduced OOV rate are exactly offset by the additional errors due to acoustic confusability. That point is the optimal vocabulary size. Figure 6 combines the slopes estimated above to arrive at a projected WER as a function of vocabulary size, for this particular task. Assuming acoustic confusability grows linearly, optimal vocabulary size is about 66K words, but the slope is very mild in the range 55KW–80KW. Assuming acoustic confusability grows logarithmically, optimal vocabulary size is in the range 80KW–110KW, but the slope is very mild starting at 70KW.

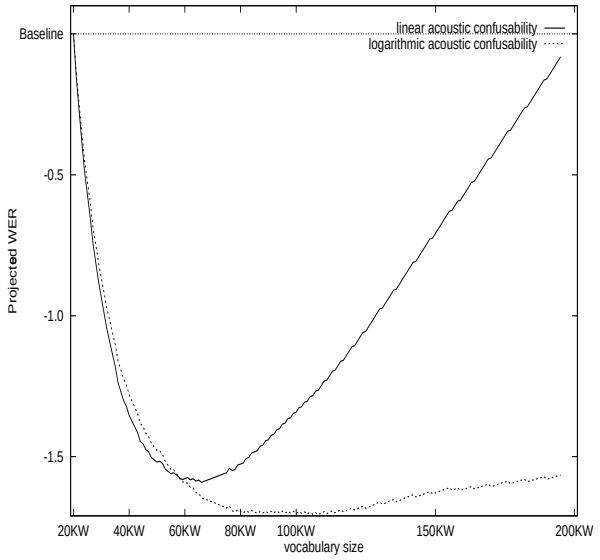

Figure 6: Projected WER based on estimated slopes for OOV errors and acoustic confusabilty. Increasing the vocabulary beyond 64KW is likely to yield negligible improvement at best.

Note though that the estimates are not very accurate. Furthermore, we do not know which of the two assumption is more correct, although it is reasonable to assume that the true answer lies somewhere in between them. I can only conclude that, for this task and with our current recognition system, increasing the vocabulary beyond the 64KW point is likely to yield negligible improvement at best.

1.3. Lexical coverage: summary

From the studies reported above I conclude that, at least in this domain:

- Lexical coverage is strongly affected by the amount of training data used to construct the lexicon, but the effect attenuates around 30MW–50MW.
- Month from which the data is drawn is insignificant.
- Source of the data is very important.
- Recency is also important: over 2 years there is a 5%–15% degradation in OOV rate. Over 4.5 years, a 11%–24% degradation..
- Best lexical coverage is achieved with only 19% of the available training data!
- With an optimized 60KW lexicon, the occurrence profile of the remaining OOV words is very unremarkable, leaving little hope for further improvement.
- Every OOV word results on average in some 1.2 word recognition errors.
- As the vocabulary grows, increased acoustic confusability is a non-negligible source of recognition errors. Since OOV rate declines at a slowing rate, there is a point of optimal vocabulary size. For this task and our current system, that point is in the range 55KW–80KW (assuming acoustic confusability grows linearly) or 80KW–110KW (assuming it grows logarithmically).

Given the last conclusion, and the limit of 64K pronunciations in our current implementation of the decoder, I settled on a vocabulary of some 59,000 words. These resulted in 64,500 pronunciations, leaving room for 1000 more to be added dynamically. The OOV rate of the 1994 eval test set with regard to this vocabulary was 0.5% (42/8186), compared to 2.4% (194/8186) relative to the official 20KW vocabulary used in the C1 run. Using the slope of 1.2 WER points for every point in OOV rate reduction, I arrive at an estimated WER reduction of 2.2% on the eval set due to the expanded and optimized vocabulary.

The extent of reduction in OOV rate due to word-order optimization depends on the vocabulary size. The larger the vocabulary, the smaller the difference, since OOV rates themselves decline rapidly. With the 59KW vocabulary, the reduction in OOV rate over the baseline (a simple top-frequency list based on the entire training corpus) was moderate (12%). But more importantly, the OOV studies revealed the dependence of lexical coverage on various aspects of the training data. This will help us determine how much (and what kind of) data we need in order to get sufficient coverage in other tasks. Moreover, the same technique can be used to study (and subsequently optimize) coverage of bigrams and trigrams. See Section for the beginning of such investigation.

1.4. Lexical coverage: analysis

In North American Business English (as defined by the 1994 NAB corpus), the least frequent among the most frequent 60K words have a frequency of about 1:7M. In optimizing a 60KW vocabulary we are thus trying to distinguish words with frequency of 1:7M from those that are slightly less frequent. To differentiate somewhat reliably between a 1:7MW word and, say, a 1:8MW word, we need to observe them enough times for the difference in their counts to be statistically reliable, i.e. we must have at least 100MW–200MW of training data. Fortunately, for constructing a decent vocabulary, it is enough that *most* such words are ranked correctly. For this, 50MW–100MW might be sufficient (since the expected difference between the counts will be 1–2). This agrees with the empirical results reported above, according to which the OOV curve improves

rapidly as more training data is used up to 50MW, and then continues to improve more slowly beyond that point.

To optimize the vocabulary for coverage of a specific time period, we must use training data from that period, or as close to it as possible. But for, say, 70MW of training data, at the DJIS wire feed data rate, we need 4 years, during which the language shifts considerably, and 60KW OOV-rate degrades by some 22% (see study of recency above). Thus we are inherently unable to fully optimize the vocabulary.

One can further generalize the last observation. Viewing language as a non-stationary stochastic source, and generalizing the word probabilities to any time-dependent linguistic phenomenon (e.g., a rise in the probability of an ngram above its static level), I arrive at the following principle:

> One can never determine accurately both the extent and the time frame of a linguistic phenomenon.

There is an inherent tradeoff between the accuracy of an estimate and the time period it is based on. More precisely, it is not the time period but the amount of training data that is the limiting factor. But since there is only a limited amount of data from each time period, the two are related by a constant.

Thus if a phenomenon is both transient and rare, we are inherently incapable of detecting it. Note that rare phenomena are not necessarily unimportant, since there may be many of them. Estimating an event as having Probability 10^{-7} rather than 10^{-6} can have a devastating effect on the log-probability, and hence recognition, of a sentence. Yet such events are commonly modelled in most existing language models and commonly encountered in test data.

The amount of LM training data available until recently was small enough that the benefit from acquiring more data dominated over the disadvantage due to language shift. But with the larger amounts of data made available recently, this is changing. With the 1994 NAB corpus of 227M words, I have already found that better vocabularies are constructed by using only a fifth of the available data. As will be seen in the next section, similar results do not yet apply to ngram lists. But with several billion words of training data, I believe they will. Language modeling is close to the point where the time-honored maxim "there's no data like more data" no longer holds.

2. Ngram Coverage and Language Model Size

In a recent work ([Chase et. al 94]) we found that recognition errors are much more likely to occur within trigrams and (especially) bigrams which have not been observed in the training data. In these cases, the language model typically relies on lower order statistics. If the bigram is missing, predictions are made based on unigram statistics, which are notoriously unreliable. Thus increased ngram coverage may translate directly into improved recognition accuracy.

But increased ngram coverage usually comes at the cost of increased memory requirements. To study the tradeoffs involved, we compared several systems on the 1994 development test. All systems used a 58KW vocabulary (different than the optimized vocabulary reported in Section) and conventional trigram backoff language models. The models differed in the amount of data they were trained on, and in their bigram and trigram cutoffs. Table summarizes the results, in decreasing order of Word Error Rate[2]. 't94' refers to the entire official 1994 NAB training corpus (227MW). '-m-n' means that bigrams occurring m or fewer times and trigrams occurring n or fewer time were excluded. The 'coverage' columns reports the rate at which the backoff language model relied on its trigram, bigram, and unigram components to produce scores (1.4% of the words were OOVs).

system	# of (M)		coverage(%)			PP	WER
	2g	3g	3g	2g	1g		
wsj93-0-0	3	7.5	57	31	11	197	
t94-1-2	6	10	63	29	6.8	156	14.7
wsj91-94-0-1	6	5	59	32	7.3	163	14.55
wsj87-94-0-1	9	8.5	63	29	6.4	153	14.35
t94-0-2	14	10	63	30	5.2	153	14.3
t94-1-1	6	18	67	25	6.8	152	14.25
t94-0-1	14	18	67	27	5.2	150	14.1*

Table 1: Ngram coverage, perplexity and Word Error Rate for LMs based on various amounts of data and different ngram cutoffs. "There's no data like more data" still holds.

A few observations:

- Given the same training data, adding bigrams or trigrams (by lowering their respective cutoffs) improves both perplexity and recognition. Interestingly, 't94-0-2' and 't94-1-1' performed similarly, even though one had 8M more bigrams while the other had 8M more trigrams.
- In the case of lexical coverage, older and less relevant training data actually hurt performance. But with ngram coverage, this does not seem to be the case. My hypothesis is that this difference is due to the much lower frequency of the ngrams (as compared to the least frequent words in the vocabulary). See discussion in Section . The largest system ('t94-0-1') performed best on the dev data, and was consequently used in our evaluaton system.

It is hard to draw further conclusions from comparing models based on different training sets. For example, 'wsj91-94-0-1' has fewer trigrams, worse test-set ngram coverage and worse test-set perplexity than 't94-1-2', and yet it performed better. Perhaps the differences in WER are not large enough to be significant. Clearly, more carefully controlled studies are called for.

References

[Roseneld 95] Ronald Rosenfeld, "The CMU Statistical Language Modeling Toolkit, and its application to the 1994 ARPA NAB Corpus", submitted to *Eurospeech 95.*

[Chase et. al 94] Lin Chase, Ron Rosenfeld, and Wayne Ward, "Error-Responsive Modifications to Speech Recognizers: Negative N-grams", in *Proc. International Conference on Spoken Language Processing,* Yokohama, Japan, September 1994.

[2]The last WER result is approximate, since it involves corrections to account for other system components.