

THE ESTIMATION OF POWERFUL LANGUAGE MODELS FROM SMALL AND LARGE CORPORA

Paul Placeway, Richard Schwartz, Pascale Fung, and Long Nguyen*

Bolt Beranek and Newman Inc.
Cambridge, MA 02138

ABSTRACT

This paper deals with the estimation of powerful statistical language models using a technique that scales from very small to very large amounts of domain-dependent data.

We begin with an improved modeling of the grammar statistics, based on a combination of the backing-off technique [6] and zero-frequency techniques [2, 9]. These are extended to be more amenable to our particular system. Our resulting technique is greatly simplified, more robust, and gives improved recognition performance than either of the previous techniques.

We then further attack the problem of robustness of a model based on a small training corpus by grouping words into obvious semantic classes. This significantly improves the robustness of the resulting statistical grammar.

We also present a technique that allows the estimation of a high-order model on modest computation resources. This allows us to run a 4-gram statistical model of a 50 million word corpus on a workstation of only modest capability and cost.

Finally, we discuss results from applying a 2-gram statistical language model integrated in the HMM search, obtaining a list of the N-Best recognition results, and rescore this list with a higher-order statistical model.

Introduction

We know that, in a real task, the importance of the language model is comparable to that of the acoustic module in determining the final performance.

Unlike most previous work on statistical language modeling, which has depended on the availability of very large text corpora, here we also deal with a special condition of severely limited training data. This is because, in general, the tasks being targeted do not currently exist as spoken

language tasks. Therefore, our only view of the task comes from limited and expensive simulations [3]. In addition, even when the task is finally implemented, the rate at which data is accumulated may be quite low. While we can construct a language model from a large general corpus and try to use it on a specific one, we know that when we do this the perplexity may increase by an order of magnitude, and the speech recognition performance is degraded, due to a mismatch between the general corpus and our specific task. Thus, it is essential to include the limited available data from the task to estimate a powerful and yet robust language model.

We use statistical language models in two different ways. The first is a bigram (order-1) grammar used internally in the Byblos HMM recognition system [1], which is used in the same way as a finite-state or a word-pair type grammar. The second use is an external method of scoring of N-Best recognition results for their language likelihood, as part of the N-Best rescoring paradigm [8]. This can be any model, but ours are generally either 3-gram or 4-gram models based on words or word classes.

A Simplification of Backing-Off

The backing-off technique [6] is very useful for robust statistical language modeling. Briefly, it says that to compute the likelihood of a novel word, a certain amount of the total probability mass for the conditioning context should be redistributed to the unobserved words, and that this redistribution should depend on the distribution of the next lower-order model.

One problem with the method, however, is that the Turing-Good method of calculating the probability of novel events, as used in [6], is not only overly complex, but also non-robust. Rather than attempt a smooth approximation of the Turing function, we used a slight modification to a much simpler technique described in [2, 9]. Rather than subtracting part of the probability mass of the conditioning context according to the occurrence statistics for that context, we add a small factor to the total observation count of the context to account for the number of occurrences of

This work was supported by the Defense Advanced Research Projects Agency and monitored by the Office of Naval Research under Contract No. N00014-89-C-0008. *Author's present address: Computer Science Department, Columbia University, New York NY 10027

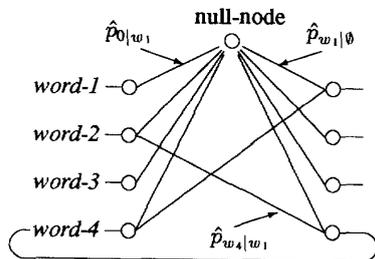


Figure 1: Example HMM Grammar Structure

previously unseen words.

Our model uses the following estimate for \hat{p} , the conditional probability of a word w in a particular context x :

$$\hat{p}_{w|x} = \frac{c_{w|x}}{n_x + r_x}$$

where $c_{w|x}$ is the number of times word w occurred in context x , n is the total number of occurrences of words in that context ($n_x = \sum_{i \in x} c_{i|x}$), and r_x is the number of *different* words that occurred in context x ($r_x = |x|$). Therefore $\hat{p}_{0|x}$, the probability of a previously unseen word occurring in that context, is given by:

$$\hat{p}_{0|x} = \frac{r_x}{n_x + r_x}$$

In recognition, if the word did not occur in this context, then $\hat{p}_{w|x}$ is calculated as the product of $\hat{p}_{0|x}$ and the conditional probability of the word occurring in the next lower-order model. This is similar to the steps taken in [6].

We have found this technique to be empirically equivalent to the Turing-Good method, and very robust in the face of abnormal data such as the DARPA 1000 word Resource Management corpus (RM), for which the Turing-Good method assumptions of the occurrence distribution are untrue. The Turing-Good model makes the assumption that more words-in-context occur once than twice, twice than three times, and so forth. For many small corpora, this is untrue, and it should not be assumed in any case.

An additional problem with the backing-off technique is that it requires an *either-or* decision based on the existence of the word in the conditioning context to determine how its probability is to be computed. For use in the HMM, this is only suitable if the grammar is fully connected. A fully connected grammar is not feasible to either compute or store, however, so we use a modified bigram grammar structure [1], as shown in figure 1, with explicit paths for the pairs of words actually observed in training,

and arcs to a single backoff node for the unigram node. In this structure, the arcs up to the backoff node are at cost $\hat{p}_{0|x}$, and the arcs down have a cost equal to the unigram probability of their target word.

The problem with this structure is that if a word pair is observed, there is not only a direct arc path but also a backing-off path from the first word to the second. Under the strict backing-off paradigm this would require elaborate decision logic to correctly implement.

To overcome this limitation, we consider all estimated probabilities to be a combination of actual observation and "getting lucky." Under this paradigm, the quantity $\hat{p}_{0|x}$ is used to smooth all probabilities with those of the next lower-order model. This is done recursively for all orders in our model. We have empirically determined that not only does this result in an improved estimate of internal bigram grammars, but also in externally computed grammars, used in the N-Best paradigm [8]. Specifically, using this refinement on the internal grammar, we observed a 10% decrease in the word error rate on the ATIS recognition task.

When we then used trigram models to rescore the N-Best hypotheses list, we observed a 40% reduction in the word error rate of utterances that were judged answerable ("class A and D") relative to the performance with the bigram grammar.

Improving Robustness

When using a small corpus, a word-based statistical grammar is still not sufficiently robust. To overcome this, a fairly common technique is grouping the words into a small number of syntactic groups. We have found that doing so overly smooths the data, resulting in an insufficiently powerful grammar.

Our solution is to group together only words in an obvious semantic class, such as the names of ships, months, digits, etc., but leaving other words in unique classes.

As a test, we compared the perplexity with three different grammars for the RM task with 100, 548, and 1000 classes respectively. In the first, words were grouped mainly on syntactic grounds, with additional classes for the short very common words. In the second, we grouped into classes only those words that obviously belonged together. (That is, we had classes for ship names, months, digits, etc.) Thus, most of the classes contained only one word. In the third grammar, there was a separate class for every word, thus resulting in a 1000-word bigram grammar. We used the backing off algorithm to smooth the probabilities for unseen bigrams. As can be seen from table 1, the 550-class grammar had a perplexity 17% lower than the 1000-word bigram, and 83% lower than the 100-class grammar.

	Number of Classes		
	100	548	1000
Training	59	19	14
Test	75	41	48

Table 1: Perplexity for three bigram class grammars measured on the training and test set.

The effective difference between the 548- and 1000-class grammars was larger than implied by the average perplexity. The standard deviation of the word entropy was one half bit higher for the 1000-class grammar, which resulted in an increase of 50% in the standard deviation of the perplexity. This indicates that the word bigram grammar frequently has unseen word pairs with very low probability, while this effect is greatly reduced in the class grammar. Thus, as expected, the class grammar is much more robust. Recognition results comparing these three grammars gave similar results, with the semantic class grammar clearly the best. An added benefit of the many-classes technique is that it is much easier to group the words into reasonable classes since only the obviously related sets of words are grouped together.

In our February 1991 evaluation on the ATIS spontaneous speech corpus, we ran two conditions: a "standard" condition which used a strict word bigram grammar, and an "augmented" condition for which the grammar included classes similar to the 550-class grammar above, and treated common strings of letters such as "T W A" as a single word. As shown in table 2 below, although the perplexities of these two grammars were very close, the recognition results for the augmented grammar were 42% better. This shows that the robustness of a grammar can greatly effect the speech recognition process.

	Test Set Grammar Perplexity	Speech Recognition Word Error
Standard	21.3	22.8
Augmented	22.2	16.1

Table 2: Comparison of grammar perplexity and actual speech recognition performance.

In order to further improve the robustness of the grammar, we investigated using cooccurrence smoothing [7] to further smooth our grammar in a manner similar to that of [5]. Unfortunately, we found that this gives only a very modest reduction in the variance of the average grammar perplexity, but almost no improvement in recognition per-

formance. We suspect that since we were running with a simple semantic class grammar, many of the cooccurring words were already grouped. Also, since this was tried on a very large corpus, we had a sufficient amount of training data, making further smoothing unnecessary.

Handling large corpora

We have developed a simple technique to deal with the implementation problems related to estimation of n -gram grammars with large vocabularies on very large training sets. The problem stems from needing to be able to store the partial probability estimates in a structure that is efficient both for searching and for adding new sequences. The natural way to do this is to use hash tables with linked lists of similarly hashed items. However, when training on 50 million words of text from the Wall Street Journal corpus (WSJ), we find 1.5 M unique 2-grams, 8 M 3-grams, and 12 M 4-grams. The virtual memory of the program quickly exceeds the 128 MB physical memory of our largest machines (the Silicon Graphics 4d/35), and the linked lists tend to be very fragmented in memory, resulting in excessive paging.

We have solved this problem in three steps. First, we distribute the training data into disjoint sets, based on a hash of the first class of each sequence. Second, we estimate the n -gram probabilities for all of the data in each set in turn. Then, the resulting probabilities can be written out in compact structures (*i.e.* arrays) that are optimized for fast searching and minimal paging, but without the capability for adding new n -grams. Third, we simply read in each of the files with estimated probabilities. Since the files contain disjoint sets of states, we do not need to merge them in any way. When we want to look up the probability of a particular n -gram, we first determine which set of probabilities to look in, based on the class of the first word of the state. The result is that we can easily store and search through the 22 million n -grams in WSJ needed for a 4-gram-based model.

We have performed experiments comparing the perplexity and accuracy with 3-grams and 4-grams for WSJ. While the perplexity with 4-grams is slightly lower (37 vs. 45), the recognition error is essentially the same (9.3% vs 9.4%). We assume that this is because the difference in perplexity is offset by a decrease in robustness. Still, it is encouraging that the 4-grams are no worse than the 3-grams. This shows that our modeling technique is robust and accurate.

N-Best Rescoring

For all of these results showing recognition results for higher-order language models, the technique used is to

decode the speech using a bigram model integrated into the HMM system, obtain an N-Best list of sentence hypotheses, then separately compute the higher-order sentence likelihood for each of these and combine this language model score with the acoustic score obtained from the HMM [8].

We have found this technique very effective. In the February 1992 DARPA evaluation on the ATIS corpus, we got an overall 20% reduction in the word error rate for all utterances. This was due to a 40% reduction in error for utterances that are considered "answerable" (classes A and D), and no reduction in error for those considered "unanswerable" (class X). That there was no improvement for the unanswerable utterances is not entirely surprising, since these sentences have statistics that are significantly different from the training, with a perplexity that is over twice that of the class A and D sentences. It is encouraging that the higher-order model did not hurt; this again shows the robustness of these techniques. Our final error rates were 6.2% for A+D and 9.4% for A+D+X.

Rescoring with the higher-order model also improved recognition accuracy for the WSJ corpus, though not as much as for ATIS. In this case the bigram recognition error was 11.4%, vs. 9.3% for the trigram, so the higher-order rescoring gave a 22% reduction in error. More recent results for the 20,000 word open-vocabulary WSJ corpus on the November 1992 evaluation are less impressive, with the bigrams scoring 16.7% vs. the trigrams at 14.8%. This is only a 13% relative gain, and is somewhat surprising, as we expected a larger improvement. This is partially due to the large number of out of vocabulary words in the open test set, which was 2-3%. On a development test set, using only utterances that were entirely in vocabulary, this same system got 16% word error using only bigrams, and 10% for the trigrams, over a 60% improvement. A second factor is the extreme length of the utterances in WSJ. We believe that there may be several regions of error in many of the utterances, but then N-Best system seems to do best if there is only one or two regions of uncertainty. We are currently working on improving these problem areas.

1. REFERENCES

- [1] Austin, S., Peterson, P., Placeway, P., Schwartz, R., Vandergrift, J., "Toward a Real-Time Spoken Language System Using Commercial Hardware," *Proc. DARPA Speech and Natural Language Workshop, Hidden Valley, PA* Morgan Kaufmann Publishers, Inc., June 1990.
- [2] Bell, T. C., J. G. Cleary, I. H. Witten, *Text Compression*. Englewood Cliffs, NJ: Prentice Hall, 1990.
- [3] Boisen, S., Ramshaw, L., Ayuso, D., Bates, M., "A Proposal for SLS Evaluation," *Proc. DARPA Speech*

and Natural Language Workshop, Cape Cod Morgan Kaufmann Publishers, Inc., Oct. 1989.

- [4] Derr, A., R. Schwartz, "A Simple Statistical Class Grammar for Measuring Speech Recognition Performance," *Proc. DARPA Speech and Natural Language Workshop, Cape Cod* Morgan Kaufmann Publishers, Inc., Oct. 1989.
- [5] Essen, U., and Steinbiss, V., "Cooccurrence Smoothing for Stochastic Language Modeling," *Proc. 1992 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, San Francisco*, vol. I, pp. 1-161-1-164, Mar. 1992
- [6] Katz, S. M., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, no. 3, pp. 400-401, Mar. 1987
- [7] Sugawara, K., Nisimura, M., Toshioka, K., Okochi, M., and Kaneko, T., "Isolated Word Recognition Using Hidden Markov Models," *Proc. 1985 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Tampa, FL*, pp. 1-4, Mar. 1985
- [8] Schwartz, R., Austin, S., Kubala, F., Makhoul, J., Nguyen, L., Placeway, P., Zavaliagos, G., "New Uses for the N-Best Sentence Hypotheses Within the Byblos Speech Recognition System," *Proc. 1992 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, San Francisco*, vol. I, pp. 1-1-1-4, Mar. 1992
- [9] Witten, I. H., T. C. Bell, "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression," *IEEE Trans. Inform. Theory*, vol. IT-37, no. 4, pp. 1085-1094, Jul. 1991