UP FROM TRIGRAMS!
THE STRUGGLE FOR IMPROVED LANGUAGE MODELS

FREDERICK JELINEK

Continuous Speech Recognition Group
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598

SUMMARY

The first experimental results in "large" vocabulary speech recognition were obtained in 1976 [Bahl 78]. They involved continuous speech reading of the so called Laser Patent Text. The test set was limited to sentences that were entirely composed of words belonging to a vocabulary of the 1000 most frequent words found in the training text. The recognizer attempted to find that sequence of words $\hat{W}$ satisfying

$$\hat{W} = \arg\max_{W} P(A \mid W)P(W) \qquad [1]$$

where A denotes the observed speech signal and W any sequence of words from the prescribed vocabulary. In the above Laser 1000 experiment, the a priori probability $P(W)$ that the speaker would utter the sequence $W = w_1 w_2 \ldots w_n$ was based on the trigram approximation

$$P(W) = P(w_1)\,P(w_2 \mid w_1)\,P(w_3 \mid w_1, w_2)P(w_4 \mid w_2, w_3) \ldots P(w_n \mid w_{n-2}, w_{n-1}) \qquad [2]$$

The probabilities $P(w_i \mid w_{i-2}, w_{i-1})$ were estimated from relative frequencies of n-grams occurring in a training text.

The surprising fact is that even now, a full 15 years later, after all the solid progress in speech recognition, the trigram model remains fundamental. Although some slightly lower perplexities [Jelinek 77] have been obtained by other means [Bahl 89], all practical alternatives involve the trigram model as a component. That this simple approach is so successful is a source of considerable irritation to me and to some of my colleagues. We have evidence that better language models are obtainable, we think we know many weaknesses of the trigram model, and yet, when we devise more or less subtle methods of improvement, we come up short.

Why should trigrams work so well? First, because they are firmly based on data, the more the better [1]. Second, because they reflect simultaneously syntax, semantics, and pragmatics of the domain in question. Finally, because European languages (not just English!) have a strong tendency toward standard word order and are thus substantially local.

Fred Damerau, an IBM linguist, participated some years ago in an experiment indicating that substantial improvement over trigrams is possible. Without benefit of any speech signal, he read the output of our recognizer and identified correctly 82 of 102 words decoded in error, while falsely judging as erroneous only 5 of a total of 2790 words read. We then asked him to replace words that he identified as wrong by those he deemed correct, whenever he was reasonably sure of the answer. He succeeded 49 times out of 102, and made the wrong guess in only 15 cases.

In another experiment, knowing the past perfectly, I attempted to guess the next word from a list (the so called *fast match list* of average size 30) of words that included the actual word and that were phonetically similar to it. My first guess was correct 91% of time while the trigram model when applied to the task was correct only 81% of the time. However, the correct word could be found in the trigram model's top two guesses 90% of the time!

The last result, while showing that improvement over trigrams is possible, indicates how hard it will be to achieve it. Another indication of the strength of trigrams is provided by the so called sentence-creation-from-a-bag task. We took 38 randomly selected sentences of length 10 words or less, and determined that permutation of their words that was most likely according to the trigram model. In 63% of the cases, the most probable permutation turned out to be the actual sentence. We carried out this experiment in conjunction with our statistical translation effort and, interestingly enough, the most probable permutation carried the same meaning as the source sentence in 86% of the cases.

[1] In 1987 Glenn Whitney ran an experiment at IBM computing the perplexity of a certain corpus as a function of the amount of text (in words) used in the computation of the language model. Here is a short table of the results:

PERPLEXITY	AMOUNT OF TEXT
85.6	1000 K
100.1	500 K
113.3	200 K
121.3	100 K

And yet, trigrams have a lot of problems. First, one would expect them to be unsuitable for highly inflected languages, like French and Italian, that should allow for a considerably freer word order than English. When a language is inflected, a much larger vocabulary is required for the same coverage of text, and a much larger corpus is necessary for extraction of word-based statistics of the same reliability. For German which compounds its content words (e.g., Wieder|gut|machung, or Reichs|sicherheits|haupt|amt) and thereby creates neologisms right and left (e.g., Mauer|specht), surely morphemes of some sort should form more appropriate building blocks of a language model than do words. But nevertheless, word trigrams serve as a backbone for all strong language models.

Trigrams do not accommodate the unquestionably dynamic character of discourse. Relative frequencies of trigrams reflect averages over the training corpus. But who of us does not believe that words cluster, that the appearance of one word will "trigger" the production of related words? That rare words like *bouquet* are to be expected when conversation turns to love or wine. Constructing a dynamic model is very important [DeMori 91, Jelinek 91] in view of the observation that words that are predicted with probability less than 10^{-4} have a 30% chance of not being recognized by Tangora.

The construction of a trigram language model is very straightforward. All that is needed is a corpus reflecting the discourse domain of the intended speech recognizer. But this is also the method's downfall: the necessity of a large sample of text, different for every category of human endeavor, be it radiology, musicology, letters to the editor, or to one's enemy or beloved. How are we going to equip our recognizers with universal and yet powerful language models easily adaptable to the changing purpose to which they will be put?

No finite vocabulary, no matter how carefully chosen, can cover fully a speaker's need. He will want to add those words that he uses routinely (names of friends, technical terms, etc) or for a particular task at hand. How can we estimate the frequencies of trigrams of an extended vocabulary?

Several of us have been struggling with the above questions and have come up with a variety of answers [Jelinek 90 and Jelinek 91] that are only partially successful, at best. These can truly be called band-aid solutions. One obstacle to progress is that the quality measure we use, perplexity, PP, is only a very crude predictor of recognizer error rate [Ferretti 90]. In fact, the formula

$$PP = \exp[\lim_{n \to \infty} (\frac{1}{n}) \log P(w_1, w_2, \dots, w_n)] \qquad [3]$$

treats words as abstract entities, and says nothing about their acoustic similarity, about the difficulty of carrying out the hypothesis search (which, given a fixed perplexity, is vastly more difficult for tasks based on a large vocabulary than on a small one), or about the peculiar strengths and weaknesses of the recognizer.

At the beginning of this article I have mentioned that trigram probabilities are estimated from relative frequencies of n-grams. The most common formula is

$$P(w_3 | w_1, w_2) = \lambda_3 f(w_3 | w_1, w_2) + \lambda_2 f(w_3 | w_2) + \lambda_1 f(w_3) \qquad [4]$$

where the weights λ_i are determined by the method of deleted interpolation [2] [Jelinek 80].

But there are many other ways of smoothing the data. Instead of being a linear combination, the estimate may be based on *backing-off* to a lower order n-gram when a higher order one is judged unreliable or is non-existent in the training text [Katz 89]. A maximum entropy approach may be used, assuring that the joint probability $P(w_1, w_2, w_3)$ satisfies marginals of whose correctness one is convinced [Brown 92]. Or additional or different components may be placed on the right-hand side of [4], perhaps based on parts of speech [Derouault 84] or on a hierarchical equivalence classification derived automatically from data [Brown 91].

Even the very idea of a relative frequency estimate of probability is open to question. Is a trigram seen k times really k times more probable than that seen only once, and is the latter infinitely more probable than one never seen? Or was the appearance of the singleton trigram only a lucky coincidence signifying very little? In fact, we can easily take the view that all trigrams seen exactly k times are equally probable, and estimate in a variety of ways the relative "worth" of these probabilities as a function of k [Jelinek 85 and Church 91].

As already indicated, none of these alternate methods of trigram probability estimation lead to any decisive breakthroughs.

There exist several approaches to language modeling that differ radically from trigrams. One of them is based on neural networks [Jain 90]. Another, TINA [Hirshman 91], comes about because the system designers of MIT's VOYAGER needed a model for speech understanding rather than for recognition. TINA and other dialogue models [Niemann 90] are both interesting and perform adequately, but only for very limited applications. Actually, it is very possible that satisfactory dialogue systems can be based on word spotting, and in that case an entirely different kind of language model would be appropriate. In fact, a fresh approach to the problem has recently been pioneered at ATT - Bell laboratories [Gorin 91].

[2] In constructing our language models, we usually add a fourth component to the righthand side of (4), namely a uniform distribution over the vocabulary (called *the floor*). We chose the interpolation weight of the floor in two ways: to minimize perplexity and to minimize the error rate. The results illustrate our contention that perplexity is not directly related to error:

Minimal perplexity case:	perplexity 139	error rate 33/1688
Minimal error rate case:	perplexity 147	error rate 24/1688

Efforts are under way to construct grammars appropriate to a relatively wide discourse domain. Progress is very slow, and performance as well as coverage are inadequate. The issue of estimation of appropriate statistical parameters of such grammars is far from resolved. It will take a while before grammar perplexities approach those of trigrams. Nobody seems yet to have formulated a credible plan to take advantage of parsers, unless it be as post processors on recognized sentences. Eventually, one would want to use the content analysis provided as a means of tracking the meandering soliloquy.

The most radical attempts at an alternate language model are based on decision trees [Bahl 89]. The idea is to ask a series of successive questions about the hypothesized history, thus determining the latter's equivalence class, and then predict the next word accordingly. The kinds of questions one asks are determined automatically from training text. In one approach, each word is specified as a string of ingeniously chosen bits, so that the history itself is a string of bits, and the questions are of the type "is the nth previous bit a 1?". The questions in another approach are "does the kth previous word belong to the ith subset?", where the character of the subsets of interest is itself determined at training time and depends on the answers to previously posed questions. The tree-based language models are very costly to construct and do achieve by themselves marginally better perplexities than do trigrams. When interpolated with trigrams, they lower the perplexity by 13%. The complexity of such hybrids has so far prevented their use.

I myself remain stubbornly committed to the search for alternatives to trigrams. Success will take a long time, but it will come. I believe that it will be based on a grammar-related approach, with substantial components of the grammar itself derived automatically from text corpora. The cost of this search can probably not be justified by the economics of speech recognition, but by the necessity to achieve text understanding, and by the right of researchers to devote themselves occasionally to the solution of intrinsically interesting questions even during the present era of sometimes senseless product competition.

REFERENCES

{Bahl 78}L.R. Bahl, J.K. Baker, P.S. Cohen, F. Jelinek, B.L. Lewis, and R.L. Mercer: Recognition of a Continuously Read Natural Corpus, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, 1978.

{Bahl 89}L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer: A tree-based statistical language model for natural language speech recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol 37, No 7, July 1989.

{Brown 91}P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer: Class-based n-gram models of natural language, submitted to Computational Linguistics, 1991.

{Brown 92}P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer: Discovering a word's parts of speech from its spelling and its neighbors, to be submitted to Computational Linguistics.

{Church 91}K.W. Church and W.A. Gale: Enhanced Good-Turing and Cal-Cat: Two new methods for estimating probabilities of English bigrams, Computers, Speech, and Language, 1991.

{de Mori 91}R. DeMori and R. Kuhn: A Cache-based natural language model for speech recognition, IEEE Transactions of Pattern Analysis and Machine Intelligence, 1991.

{Derouault 84}A-M. Derouault and F. Jelinek, Modele probabiliste d'un langage en reconnaissance de la parole, Annales des Telecommunications, tome 39, no 3-4, 1984.

{Ferretti 90}M. Ferretti, G. Maltese, and S. Scarci: Measuring information provided by language model and acoustic model in probabilistic speech recognition: theory and experimental results, Speech Communication 9, pp531-539, 1990.

{Gorin 91}A.L. Gorin, S.E. Levinson, and A.N. Gertner: Adaptive acquisition of spoken language, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toronto, 1991.

{Hirshman 91}L. Hirshman, S. Seneff, D. Goodine, and M. Phillips: Integrating syntax and semantics into spoken language understanding, Fourth DARPA Speech and Natural Language Workshop, Asilomar, CA, 1991

{Jain 90}A.N. Jain and A.H. Waibel: Robust Connectionist Parsing of Spoken Language, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Albaquerque, 1990.

{Jelinek 77}F. Jelinek, R.L. Mercer, L.R. Bahl, and J.K. Baker, Perplexity - A Measure of Difficulty of Speech Recognition Tasks, 94th Meeting of the Acoustic Society of America, Miami Beach, December 1977.

{Jelinek 80}F. Jelinek and R.L. Mercer: Interpolated Estimation of Markov Source Parameters from Sparse Data, Pattern Recognition in Practice, E.S. Geltsema and L.N. Kanal (Eds.), North Holland, Amsterdam.

{Jelinek 85}F. Jelinek and R.L. Mercer: Probability distribution estimation from sparse data, IBM Technical Disclosure Bulletin 28 (6): 2591-2594, 1985.

{Jelinek 90}F. Jelinek, R.L. Mercer, and S. Roukos: Classifying words for Improved Statistical Language Models, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Albaquerque, 1990.

{Jelinek 91}F. Jelinek, B. Merialdo, S.Roukos, and M. Strauss: A Dynamic Language Model for Speech Recognition, Fourth DARPA Speech and Natural Language Workshop, Asilomar, CA, 1991.

{Katz 89}S. Katz, Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-34 (3), 1989.

{Niemann 90}H. Niemann: The interaction of word recognition and linguistic processing in speech understanding, NATO Advanced Study Institute, Speech Recognition and Understanding, Cetraro, 1990.