Is N-Best Dead?

Long Nguyen, Richard Schwartz, Ying Zhao, George Zavaliagkos†

BBN Systems and Technologies
Cambridge, MA 02138
†Northeastern University

ABSTRACT

We developed a faster search algorithm that avoids the use of the N-Best paradigm until after more powerful knowledge sources have been used. We found, however, that there was little or no decrease in word errors. We then showed that the use of the N-Best paradigm is still essential for the use of still more powerful knowledge sources, and for several other purposes that are outlined in the paper.

1. INTRODUCTION

The N-Best Paradigm [1] was introduced originally as a means for integrating the speech recognition and language understanding components of a spoken language system. Since then, we have generalized its use for integrating into the recognition search other expensive knowledge sources (such as higher-order n-gram language models, between-word co-articulation models, and segmental models) without increasing the search space [2]. The basic idea is that we use inexpensive knowledge sources to find N alternative sentence hypotheses. Then we rescore each of these hypotheses with the more expensive and more accurate knowledge sources in order to determine the most likely utterance. The N-Best Paradigm specifically, and multi-pass search algorithms in general, are now used widely by the speech recognition research community.

Besides its use as an efficient search strategy, the N-Best Paradigm has been used extensively in several other ways [2]. Its simplicity has made it ideal as a means for cooperation between research sites. For example, we regularly send the N-Best lists of alternatives to research sites that do not have an advanced speech recognition capability (e.g., Paramax and NYU) in order that they can apply their own linguistic components for understanding or for research into alternative language modeling techniques.

Another related use of the N-Best lists is for evaluation of alternative knowledge sources. New knowledge sources can be evaluated without having to integrate them into the search strategy. For example, we can determine whether a new prosodic module or linguistic knowledge source reduces the error rate when used to reorder the N-Best list. This is particularly important for knowledge sources that are not easily formulated in a left-to-right incremental manner.

Finally, we have presented techniques for optimizing the weights for different knowledge sources, and for discriminative training [2].

In this paper we attempt to determine whether the N-Best Paradigm results in substantial search errors. If it does, then its use for the other purposes mentioned above would also be questionable. First we describe briefly how we used the N-Best paradigm in previous versions of BYBLOS. Then, we describe our attempts to avoid the errors that might be a result of using the N-Best paradigm. Finally, we argue that there will always be cases where the N-Best paradigm will make it possible to use some knowledge sources that would likely never be used otherwise.

2. 3-PASS N-BEST SEARCH STRATEGY

The BYBLOS system has been described previously (e.g., [3]). We reiterate here the use of the N-Best Paradigm in that system.

The decoder used a 3-pass search strategy. The strategy used a forward pass followed by a backward Word-Dependent N-Best search algorithm [4] using a bigram language model, within-word triphone models, and top-1 (discrete VQ) densities. The N-Best hypotheses were then rescored using cross-word triphone context models, top-5 mixture densities, and trigram language model.

Typically, the backward Word-Dependent N-Best pass requires about half the time required by the forward pass. Rescoring each alternative sentence hypothesis individually with cross-word triphone models only requires about 0.2 seconds per hypothesis. And rescoring the text of the hypotheses with a high-order n-gram language model [5] requires essentially no time.

3. ADMISSIBILITY

It has often been asserted that the N-Best paradigm is inadmissible because when the initial N-Best list is created using weaker knowledge sources, then the answer that would have had the highest score using the stronger knowledge sources might not be within the list of alternatives, and therefore never have a chance to be rescored. This would be especially likely when the error rate is high and the utterances

are long, since the number of alternative sentences needed to include the correct answer would grow exponentially with the length of the utterance.

The knowledge sources (e.g., cross-word triphones and trigram language models) used for rescoring in the 3-Pass N-Best strategy described above were much more powerful than the original knowledge sources (e.g, within-word triphones and bigram language models) in that they frequently reduced the error rate by half. However, we had assured ourselves that, at least for moderate-size problems (like ATIS with 2,000 words or WSJ with 5,000 words), there were few if any additional errors caused by the correct answer not being included in the N-Best list.

However, after the November 1992 DARPA Continuous Speech Recognition (CSR) evaluations, we were concerned that we might be losing some performance as a result of our use of the 3-Pass N-Best strategy (rescoring with cross-word triphones, top-5 mixture densities, and trigram language models) on the 20,000 words WSJ test. This was because there were many sentences for which the correct answer was not in the N-Best hypotheses although it had a higher total score (when including the trigram language model and cross-word triphones) than any sentence hypothesis in the N-Best list. We felt that this was due to the higher word error rate that resulted from recognition with a large vocabulary of 20,000 words, and the long utterances found in the Wall Street Journal (WSJ) corpus.

Therefore, this year we implemented a more complicated search strategy similar to the Progressive-Search strategy suggested by Murveit [6] in which we use the initial passes to create a lattice of alternative hypotheses, which can then be rescored. The advantage of this approach is that a lattice with a small number of alternatives at each point can represent a very large number of alternative sentence hypotheses. In addition, rescoring the lattice of alternatives is computationally less expensive than rescoring a large explicit list of sentence alternatives. This also avoids the rather large intermediate storage required to store the N-Best hypotheses.

3.1. 4-Pass Lattice Search Algorithm

In this section we describe a 4-Pass Lattice Search algorithm that avoids the early use of the N-Best.

1. The time-synchronous beam search algorithm with a vocabulary of 20,000 words and a bigram language model typically requires substantial computation on today's workstations. Therefore, we make extensive use of the Normalized Forward Backward Search algorithm [8] to reduce computation. We perform a first pass using a fast match technique whose sole purpose is to find and save high scoring word ends. Because this model is approximate, it can run considerably faster than the usual beam search. And because the later passes will be more accurate, the first pass need not be as accurate.

2. A second pass, time-synchronous beam search, using a bigram language model, within-word triphones, and (top-1 VQ) discrete models runs backward. This pass is sped up by several orders of magnitude by using the Normalized Forward Backward pruning on the word-ending scores produced by the first pass. We save the beginning times and scores ($\beta^t_{w_i}$) of all words found. This pass requires much less time than the first pass.

3. A third pass identical to the second pass runs forward, using the Normalized Forward Backward pruning on the word-beginning scores produced by the second pass. Similar to the second pass, we save the ending times and scores ($\alpha^t_{w_j}$) of all words found (constrained by the second pass).

4. We use the beginning ($\beta^t_{w_i}$) and ending ($\alpha^t_{w_j}$) scores from passes 2 and 3 to determine possible word-juncture times. Specifically, if the forward-backward score for a particular pair of words is within a threshold of the total score for the utterance, then this word-pair is used. That is if

$$\alpha^t_{w_j}\ Pr(w_i|w_j)\ \beta^t_{w_i} > \lambda$$

where $Pr(w_i|w_j)$ is the probability of w_j followed by w_i, and λ is the threshold (which can be a function of either α or β).

Adjacent word-junctures are merged. Having found a word-pair, we look for the next word-juncture where this second word is the first word of the next pair. The result is a lattice of word hypotheses. If the range of beginning and ending times for a single word overlap, then we create a loop for that word.

The word lattice (which is really just a small finite-state language model) is then expanded to allow for maintaining separate scores for trigram language models and cross-word triphones. This entails copying each word in the context of each preceding word, and replacing the triphones on either side of the word junctures with the appropriate cross-word triphones. Thus, each word in the lattice represents a particular instance of that word in the context of some particular other word. The transition probabilities in the lattice are the probability of the next word given the previous two words – trigram probabilities.

We perform a fourth pass in the backward direction using this expanded language model. The result is the most likely hypothesis including cross-word and trigram knowledge sources.

However, we are not done at this point, because we may want to apply more powerful, but more expensive, knowledge sources. We generate the N-Best alternative hypotheses out of the search on this lattice. The Word-Dependent N-Best algorithm [4] requires that we keep separate scores

at each state depending on the previous word, because the boundary between two words clearly depends on those two words. But the words in the lattice are only defined in the context of the neighboring word. Thus, by keeping the scores of all of the ending word hypotheses, we can recover the N-Best alternatives. However, in contrast to its previous use, these N-Best answers have been computed including the more powerful knowledge sources of cross-word coarticulation models and trigram language models.

3.2. Experimental Results

We performed an experiment in which we compared the recognition accuracy of this 4-Pass Lattice approach with the previous 3-Pass N-Best approach. In both cases, the initial search (in order to create the lattice or to find the N-Best sentence hypotheses) used only a bigram language model and within-word coarticulation models with top1-VQ discrete densities, while the final search (on the lattice) or rescoring (the N-Best) used a trigram language model and between-word coarticulation models with top-5 mixture densities.

Initially, we were surprised to find that the accuracy using the lattice was actually slightly worse than that of the original N-Best method. Then, we realized that this was due to the larger number of alternatives. A lattice with an average depth (the average number of branches out of a word-end node) of 10 for a sentence of 20 words can be thought of as an N-Best list with 10^{20} hypotheses. When we had previously found that, in the 3-Pass N-Best approach, the correct utterance might have a higher score than the answers in the top 100 best hypotheses, there were also countless other incorrect hypotheses, in the 4-Pass Lattice approach, that also had higher scores than the answers in the original N-Best. The search on the lattice often found one of these other incorrect answers.

We alleviated this problem by optimizing (automatically) the weights (for trigram language model, word insertion penalty and phone insertion penalty) using the N-Best alternative hypotheses found after the lattice search. These new weights were then used to search the lattice again. Finally, we were able to obtain 5% fewer word errors using the 4-Pass Lattice strategy than when using the 3-Pass N-Best approach. This was a much smaller reduction in error than we had hoped for. Apparently the reduced search errors were largely offset by the larger search space on the lattice.

It would appear, therefore, that the doom and gloom predictions for N-Best are unfounded so far, at least for the 20,000 WSJ task. In fact, the N-Best paradigm continues to offer advantages not available otherwise, as mentioned below.

4. CURRENT USES FOR N-BEST

While it is possible to expand a lattice of alternatives for rescoring with trigram language models, there are still many knowledge sources that are too expensive to use this way. For example, for the November of 1993 evaluations, we included a model of whole segments (Segmental Neural Network [10]). And Boston University also rescored our N-Best hypotheses with a similarly motivated Stochastic Segment Model [9]. Both of these models are much more expensive than HMM models due to their constrained slope features and global dependence. Either of these models reduce the word error rate by about 10% in combination with the HMM scores. We also experimented with a more complex HMM topology for a phoneme that includes thirteen states instead of the usual three or five states. While this model could have been integrated directly, it was much easier and faster to simply rescore the N-Best hypotheses with this larger model. The resulting small reduction in error rate would not have been worth the larger computation and storage associated with using it in the original search, if not to mention the time of implementation to integrate these models into the search.

Also for the 1993 evaluations on the ATIS domain, we found that we could reduce the word error rate by 8% by rescoring the N-Best hypotheses with a four-gram class language model. Again, expanding the word lattice for a four-gram language model would have been possible, but would have resulted in a huge lattice with the same word replicated many times. But rescoring the N-Best hypotheses with four-grams required almost no computation and did not require rerunning the recognition.

There is a tremendous advantage in being able to define any scoring function without having to get involved with the details of a general search strategy since only one linear sequence of words need be scored at one time.

In combining these various experimental knowledge sources, it is important that they be weighted appropriately, or else there may be no gain, or even a loss. Optimizing the weights for several knowledge sources on a development test set of several hundred sentences can be accomplished in seconds or minutes on the N-Best hypotheses rather than days by explicit experimentation.

And of course, we still use the N-Best paradigm to combine the speech recognition with the language understanding component. It would be infeasible to use the entire constrained space defined by the understanding model in the speech recognition search. But it is a trivial matter to provide several (5 to 10) alternatives to the understanding component for its choice. Again, in this year as in the past, we also provided the N-Best alternatives output from our speech recognition system to the language understanding group at Paramax. This simple text-based interface makes arbitrary integration simple. The integration between two sites across the ARPA network was quite straightforward.

5. CONCLUSIONS

We developed a search strategy similar in spirit to the Progressive Search technique [6] that allows us to incorporate cross-word triphones and trigram language models directly within the search. The resulting search, although using many passes is considerably faster than our previous strategy. However, we found only marginal improvements in the accuracy, indicating that there were not really many search errors incurred using the original 3-Pass N-Best strategy.

Despite the ability to integrate some knowledge sources directly into the search, we still use the N-Best Paradigm in all the ways that we used it previously, including integrating more expensive knowledge sources, cooperation with other research sites, rapid testing of new knowledge sources, and automatic optimization of recognition parameters.

6. 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. Schwartz, R., and Y. Chow, "The N-Best Algorithm: An Efficient Procedure for Finding the Top N Sentence Hypotheses", *Proc. of DARPA Speech and Natural Language Workshop*, Cape Cod, MA, Oct. 1989, pp. 199-202. Also *Proc. of IEEE ICASSP-90*, Albuquerque, NM, Apr. 1990, S2.12, pp. 81-84.
2. Schwartz, R., S. Austin, F. Kubala, and J. Makhoul, "New Uses for the N-Best Sentence Hypotheses Within the BYBLOS Speech Recognition System", *Proc. of IEEE ICASSP-92*, San Francisco, CA, March 1992, pp. I.1-I.4.
3. 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.
4. Schwartz, R. and S. Austin, "A Comparison Of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses", *Proc. of IEEE ICASSP-91*, Toronto, Canada, pp. 701-704.
5. Placeway, P., R. Schwartz, P. Fung, and L. Nguyen, "The Estimation of Powerful Language Models from Small and Large Corpora", *Proc. of IEEE ICASSP-93*, Minneapolis, MN, Apr. 1993, Vol. II, pp. 33-36.
6. Murveit, H., J. Butzberger, V. Digalakis, M. Weintraub, "Progressive-Search Algorithms for Large Vocabulary Speech Recognition", *Proc. of ARPA Human Language Technology Workshop*, Princeton, NJ, Mar. 1993, pp. 87-90.
7. Austin, S., Schwartz, R., and P. Placeway, "The Forward-Backward Search Algorithm", *Proc. of IEEE ICASSP-91*, Toronto, Canada, pp. 697-700.
8. Nguyen, L., R. Schwartz, F. Kubala, and P. Placeway, "Search Algorithms for Software-Only Real-Time Recognition with Very Large Vocabularies", *Proc. of ARPA Human Language Technology Workshop*, Princeton, NJ, Mar. 1993, pp. 91-95.
9. Ostendorf, M., A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses", *Proc. of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, Feb. 1991, pp. 83-87.
10. Zavaliagkos, G., S. Austin, J. Makhoul, R. Schwartz, "A Hybrid Continuous Speech Recognition System Using Segmental Neural Nets With Hidden Markov Models", *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific Publishing Company 1993, Vol. 7, No. 4, pp. 949-963.