Using a Broad Coverage Precision Grammar as a Language Model for ASR

Woodley Packard

March 11, 2013

1 Introduction

Automatic speech recognition is a difficult task typically divided into two components: an acoustic model, which accounts for the relationship between the incoming audio signal and the basic logical units of speech, and a language model, which accounts for the fact that some sequences of sounds or words are much more likely than others, allowing the system to infer information that is missing in the actual signal.

The nature of the search problem puts strict constraints on the types of language models that can be used during the early stages of speech recognition. Bigram models are the most common, as they can be processed with modest CPU and memory requirements relative to more complicated models. Even the best of recognizers make plenty of errors, so the result is typically either an implicit or explicit representation of the most likely N hypotheses, with N ranging from perhaps a hundred for explicit representations to millions for implicit representations (lattices).

2 Related Work

The use of more accurate language models has proven a profitable area of research. Word lattices can be efficiently rescored with higher-degree n-gram models, frequently leading to significant reductions in word error rate (WER).

Another approach is to use a more linguistically motivated mechanism to reject nonsensical hypotheses. Georgescul et al. (2008) report a $10-20\%$ relative reduction in WER by using a support-vector machine built on linguistically motivated features to rerank an N-best hypothesis list.

McNeill et al. (2006) achieve a more modest relative WER reduction of a few percent by rescoring with a probabilistic context free grammar (PCFG).

Beutler et al. (2005) use a precision grammar for German based on the Head-driven Phrase Structure Grammar (HPSG) formalism to rescore a 20-best list *after* applying 4-gram reranking to lattices generated by a bi-gram decoder, and report a 27% relative WER reduction beyond the 4-gram reranker. In his PhD thesis, Kaufmann (2009) reports a 10% relative WER reduction, again from using an HPSG to rerank German ASR results.

3 Experimental Setup

I attacked the problem of N-best reranking for English using an existing HPSG grammar.

The English Resource Grammar (ERG)

The ERG (Flickinger, 2000) is a broad-coverage precision grammar of English. It achieves roughly 90% coverage on unseen open-domain edited[1] English text. It contains limited support for some of the types of disfluencies found in spoken English, but in general it better describes fluent English. For complex sentences, the ERG usually produces multiple candidate analyses – sometimes thousands or millions. A discriminative parse ranking system is included with the grammar, which assigns a numerical score to each candidate parse.

The TIMIT Corpus

The TIMIT corpus is a collection of relatively high-quality recordings of "phonetically rich" utterances, with several hundred speakers representing eight dialects Garofolo et al. (1993). The utterances are highly grammatical (93% coverage with the ERG). This works in my favor, since the grammar can distinguish correct results from agrammatical ones. However, results could be dramatically different in a less fluent domain. Parsing with the ERG can be a time consuming process, but since the average sentence length in TIMIT is only about 10 words, this did not prove to be a large problem[2]. The TIMIT corpus is divided into pre-specified "train" and "test" sections; I evaluated only on the "test" section.

[1] Not text messages, for example!

[2] I used the ACE parsing system: http://sweaglesw.org/linguistics/ace/

Baseline Model

For a baseline model, I trained a bigram language model on the TIMIT corpus transcriptions. Initially, I used a model trained on the TIMIT "train" section, but accuracy was so low that the correct transcription rarely appeared in the 100-best list. WER was in the 40's. I also tried using a generic bigram model that I found online, trained on the Gigaword corpus, but performance was even worse with this model (WER was in the 60's). In order for the grammar-based reranker to be able to offer any insight, I needed a system where the correct transcription usually was present in the 100-best list.

To be able to move on with my experiments, I decided to use a bigram model trained on the TIMIT "test" section. The actual accuracies that I obtain with this model are of course biased and cannot be compared to any self-respecting literature. However, I think it is fairly reasonable to expect that improvements *beyond* this artificially good baseline could generalize. This model performed well enough that the correct transcription usually appeared somewhere in the 100-best list.

I used acoustic models trained by Keith Vertanen[3]. I used HTK's "HVite" decoder tool to generate N-best lists, and the "HResults" error-rate scoring tool for all of my evaluations.

4 Results

Baseline and Oracle

The baseline model scored a word error rate (WER) of 8.41; this corresponds to picking the correct transcription of the whole sentence 63.57% of the time.

I also evaluated an "oracle reranker" model, which picks the correct transcription if it is present on the 100-best list, and otherwise picks the baseline recognizer's top scoring hypothesis. The correct transcription was present in the 100-best list for 76.37% of the items, yielding a WER of 6.22.

Using the ERG as a Filter

My first experiment was to simply reject all ungrammatical hypotheses. I attempted to parse each of the top 100 recognizer hypotheses for each test item. If a parse was found for any of them, the one that received the top

[3]Accessed from http://keithv.com/software/htk/us/ on March 7th, 2013. The models are trained on the WSJ SI-284 dataset.

recognizer score from the grammatical subset was selected. Otherwise, the original top hypothesis was chosen. This system had a WER of 7.92, which is a relative reduction of 5.8% over the baseline.

Reranking with Disambiguation Scores

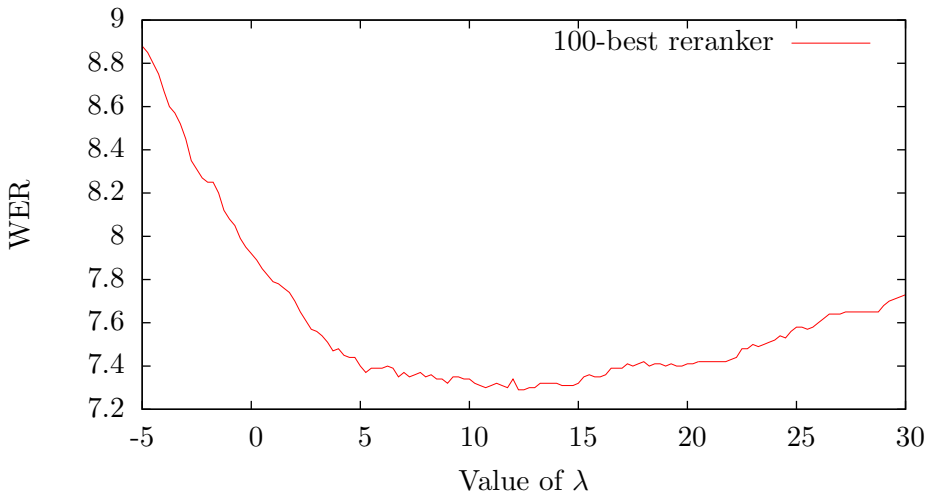
The ERG does not provide a calibrated judgement of the *degree* of fluency of inputs; rather, it is only capable of issuing a "yes" or "no" judgement as to the grammaticality of a string. However, the discriminative parse ranking module does produce a numerical score for each candidate analysis. Strictly speaking, it is only meaningful to compare these scores between multiple candidate parses of the same string. However, I decided to see whether these scores could be used as a proxy for a fluency measure.

The modified system works just like the previous one if no grammatical hypotheses are found. However, in the case where multiple grammatical hypotheses are found, it reorders them and selects a new best candidate according to the formula:

$$\text{reranked-score}(h) \;=\; \text{recognizer-score}(h) + \lambda \cdot \text{grammar-score}(h)$$

$$h_{\text{best}} = \text{argmax}_h \;\; \text{reranked-score}(h)$$

where λ is a weight parameter. The following figure shows the WER of the resulting system as a function of λ:



The case $\lambda = 0$ corresponds to the previous system. The best-performing system corresponds roughly to $\lambda = 12.5$, with a WER of 7.29. This repre-

sents a 13.2% relative reduction from the baseline rate of 8.41. Note that this is more than halfway to the oracle score of 6.22.

5 Conclusion and Outlook

I presented a method for using a precision grammar to rerank recognition results, achieving a 13.2% relative WER reduction. Since the system I evaluated uses a first-pass language model that is trained on the testing data, it is hard to predict what the result of the reranking system would be when applied to an unbiased baseline system with similar performance. Furthermore, the λ parameter was estimated on the testing data rather than on held-out data. However, to me it seems plausible that a significant reduction in WER would remain.

References

Beutler, R., Kaufmann, T., and Pfister, B. (2005). Integrating a non-probabilistic grammar into large vocabulary continuous speech recognition. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 104–109. IEEE.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 932:27403.

Georgescul, M., Rayner, M., Bouillon, P., and Tsourakis, N. (2008). Discriminative learning using linguistic features to rescore n-best speech hypotheses. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 97–100. IEEE.

Kaufmann, T. (2009). *A rule-based language model for speech recognition*. PhD thesis, Swiss Federal Institute of Technology Zurich.

McNeill, W. P., Kahn, J. G., Hillard, D. L., and Ostendorf, M. (2006). Parse structure and segmentation for improving speech recognition. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 90–93. IEEE.