A Simple Statistical Class Grammar for Measuring Speech Recognition Performance

Alan Derr
Richard Schwartz

BBN Systems and Technologies Corporation
Cambridge, MA 02138

ABSTRACT

In this paper we will discuss our development of a new grammar that is to be used for evaluation of speech recognition systems. The grammar is a statistical first-order class grammar and has been developed for for two different task domains (the DARPA 1000-word Resource Management domain and a 2000-word personnel database domain). We will first motivate the development of this grammar, next describe the grammar and its development, and finally present results and conclusions.

1 MOTIVATION

In recent DARPA speech community-wide recognition system evaluations, the recognition systems have been tested using two grammatical conditions: no grammar (or null grammar), and the word-pair grammar. These grammars suffer from several inadequacies.

The null grammar simply forces the recognition system to partition the input speech into whole-word units without using any knowledge of the language to place restrictions on the possible sequences of words that are allowed. As a result, the "no grammar" test condition provides only a worst case recognition test point for the evaluation of recognition systems.

The word-pair grammar, on the other hand, was derived from the sentence patterns that were used to generate the 2800 sentences in the Resource Management corpus. Only pairs of words that *could* occur in a sentence generated by the patterns was allowed in the word-pair grammar. On the average, each word in the vocabulary can be followed by about 60 words. No probabilities are assigned to the different words (just 0 or 1). As a result, the recognition rate is artificially high, since many reasonable word sequences are disallowed. At the same time, if a real sentence has one of these disallowed word-pairs, it could not be recognized correctly.

As a result of the unrealistic restrictions imposed by the word-pair grammar, the recognition performance of systems using this grammar is too high to allow reliable measurement of system improvements without resorting to the use of very large evaluation test sets. Creating new sentences for the test set becomes a problem, since there is a danger that new sentences that are within the task domain may not be parsed (allowed) by the word-pair grammar.

We desired a grammar that would overcome the deficiencies of the null and word-pair grammars, while at the same time providing several additional benefits. We wanted the new grammar to capture statistics that are representative of the real data in the task domain while, at the same time, providing full coverage (i.e., allowing all sentences that are possible within the task domain to be parsed by the grammar). We also wanted the grammar to be "tunable" to some degree, by allowing its perplexity to be adjusted. Increasing the grammar's perplexity will allow us to simulate a recognition system's performance with a more difficult (e.g., larger) task domain. For this reason we used several approximations in the method for estimating the grammar that caused the grammar to have higher perplexity. Finally, we wanted a grammar that would allow us to change task domains with relative ease.

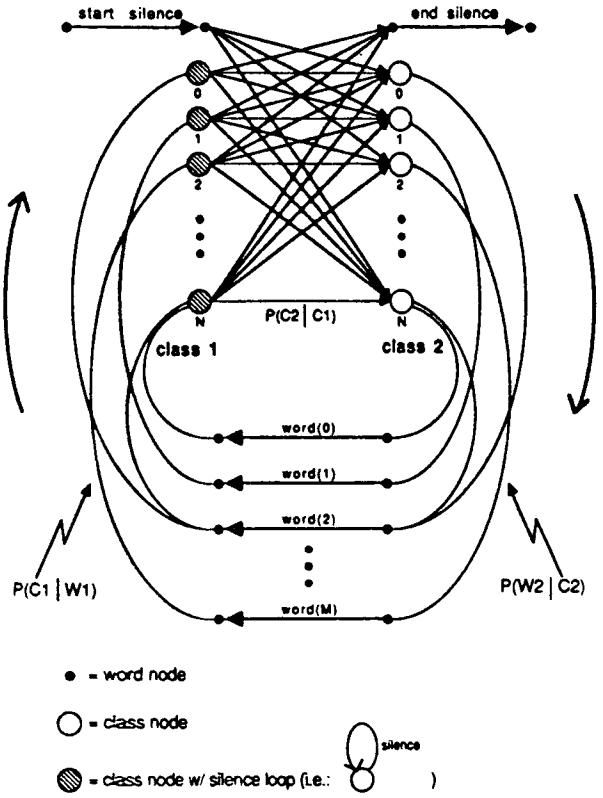
2 DESCRIPTION

The grammar that we developed is a statistical first-order class grammar in which the probability of a word ($W1$) being followed by another word ($W2$) is given by:

$$P(W2|W1) = \sum_{paths} P(C1|W1)P(C2|C1)P(W2|C2)$$

Where $C1$ is each of the classes to which $W1$ belongs, and $C2$ is each of the classes to which $W2$ be-

longs. Since each of $W1$ and $W2$ may belong to multiple classes, the summation is over all possible paths from $W1$ to $W2$. This is represented graphically below:



Note that, in the diagram, the silence at the beginning of a sentence ("start silence") and the silence at the end of a sentence ("end silence") are simply special cases of $W1$ and $W2$, respectively, where each is in a separate class. The "class node w/ silence loop" indicates that a silence may be inserted between each word.

In our work to date, we have made two simplifying assumptions. The conditional probability $P(C1|W1)$ is approximated by:

$$P(C1|W1) = \begin{cases} N_{W1 \ni C1}^{-1} & \text{for W1 in C1} \\ 0 & \text{otherwise} \end{cases}$$

Where $N_{W1 \ni C1}$ is the number of classes of which word $W1$ is a member. (For example, if a word is a member of two classes, $P(C1|W1)$ will be 0.5 for each of those classes and 0.0 for all other classes.) A similar approximation is made for P(W2—C2), where:

$$P(W2|C2) = \begin{cases} N_{W2 \in C2}^{-1} & \text{for W2 in C2} \\ 0 & \text{otherwise} \end{cases}$$

Where $N_{W2 \in C2}$ is the number of words in class $C2$. The probabilities $P(C1|W1)$ and $P(W2|C2)$ are fixed and not changed during the training of the grammar. With this simplification, the only term that must be estimated during the training of the grammar is $P(C2|C1)$, the class-to-class transition probabilities.

3 GRAMMAR TRAINING

To train the grammar, we began by assigning class(es) to each word in the vocabulary. A word may be assigned to multiple classes. For example, the word "SEAWOLF" is assigned to one class: ship-name. On the other hand, the word "DISPLAY" is assigned to three classes: command-verb, adjective, and noun. Once the words are assigned to appropriate classes, the statistics of the grammar were counted directly from the training data by counting the number of transitions from each class to each other class. These counts were then padded slightly (to account for unobserved class-to-class transitions) to allow the grammar to parse sentences containing unobserved class transitions. Finally, the grammar was tested on a test set to measure its perplexity.

4 GRAMMAR PERFORMANCE

Below is a summary some of the characteristics of the statistical first-order class grammar with 99 classes for the DARPA 1000-word Resource Management task domain, with null and word-pair grammar characteristics given for comparison.

Grammar	Coverage	Perplexity Train	Perplexity Test	Recog. error
Null	100%	992	992	12.6%
Word-pair	80%	60	NA	2.5%
Stat. class	100%	72	77	5.9%

Table 1: Grammar Performance Comparison

The class grammar figures given here are based on a grammar trained using all 2800 sentences available for

the 1000 word DARPA resource management task domain. The training set perplexity is computed over the entire training set and the test set perplexity is computed over the 300 sentences used for the May 1988 standard system evaluation. The word-pair coverage and perplexity are approximate theoretical figures assuming an independent test set. The test set perplexity for the word-pair is degenerate, since, if a single sentence doesn't parse, the perplexity becomes infinite.

In informal tests, we were able to "tune" the perplexity of the grammar by adjusting the number of classes into which the words are categorized. On a fixed test set of 100 sentences and the full training set of 2800 sentences, the perplexity varied from 203 (with 50 classes) to 62 (with 168 classes).

We have obtained some preliminary results for a statistical class grammar designed for a 2170 word personnel database access task domain. The grammar uses 637 classes (1 to 5 per word) and is trained using 750 sentences. The perplexity of this grammar on an independent test set of 200 sentences was measured to be 89.4. The perplexity measured on the training set was 46.1. We haven't yet performed a full set of recognition experiments using this grammar.

5 CONCLUSIONS

We have described the development of a statistical first-order class grammar. The structure of this grammar allows for relatively easy development of a new grammar for a new task domain. The grammar provides for full coverage of the task domain, even if all possible class sequences are not observable in the data used to train the grammar probabilities. It also provides a method for adjusting the perplexity of the grammar by varying the number of classes in the grammar.

We recommend that this grammar should be made another standard grammar for the DARPA speech community. We believe that this grammar could extend the life of the Resource Management task by decreasing the recognition system's performance while still placing restraints on the possible sequencing of words in a meaningful way. Decreasing the recognition performance will allow the (statistically significant) measurement of small system improvements without needing to increase the size of the evaluation test. We also recommend that this grammar replace the word-pair for official system evaluations.

Acknowledgements

The work reported here was supported by the Advanced Research Projects Agency and was monitored by the Office of Naval Research under N00014–85–C–0279. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.