Figures of Merit for Best-First Probabilistic Chart Parsing

Sharon A. Caraballo and Eugene Charniak

Department of Computer Science
Brown University
Providence, Rhode Island 02912

CS-96-12
April 1996

Figures of Merit for Best-First Probabilistic Chart Parsing

Sharon A. Caraballo
sc@cs.brown.edu

Eugene Charniak
ec@cs.brown.edu

February 16, 1996

Abstract

Best-first parsing methods for natural language try to parse efficiently by considering the most likely constituents first. Some figure of merit is needed by which to compare the likelihood of constituents, and the choice of this figure has a substantial impact on the efficiency of the parser. While several parsers described in the literature have used such techniques, there is no published data on their efficacy, much less attempts to judge their relative merits. We propose and evaluate several figures of merit for best-first parsing.

1 Introduction

Chart parsing is a commonly-used algorithm for parsing natural language texts. The chart is a data structure which contains all of the constituents which may occur in the sentence being parsed. At any point in the algorithm, there exist constituents which have been proposed but not actually included in a parse. These proposed constituents are stored in a data structure called the keylist. When a constituent is removed from the keylist, the system considers how this constituent can be used to extend its current structural hypothesis. In general this can lead to the creation of new, more encompassing constituents which themselves are then added to the keylist. When we are finished processing one constituent, a new one is chosen to be removed from the keylist, and so on. Traditionally, the keylist is represented as a stack, so that the last item added to the keylist is the next one removed.

Best-first chart parsing is a variation of chart parsing which attempts to find the most likely parses first, by adding constituents to the chart in order of the likelihood that they will appear in a correct parse, rather than simply popping constituents off of a stack. Some figure of merit is assigned to potential constituents, and the constituent maximizing this value is the next to be added to the chart.

In best-first probabilistic chart parsing a probabilistic measure is used. In this paper we consider probabilities primarily based on probabilistic context-free grammars, though in principle other, more complicated schemes could be used.

Ideally, we would like to use as our figure of merit the conditional probability of that constituent, given the entire sentence, in order to choose a constituent that not only appears likely in isolation, but maximizes the likelihood of the sentence as a whole; that is, we would like to pick the constituent that maximizes the following quantity:

$$p(N^i_{j,k}|t_{0,n})$$

where $t_{0,n}$ is the sequence of the n tags in the sentence (numbered $t_0, \ldots, t_{n-1}$), and $N^i_{j,k}$ is a nonterminal of type i covering terms $t_j \ldots t_{k-1}$. However, we cannot calculate this quantity, since in order to do so, we would need to completely parse the sentence. In this paper, we examine the performance of several proposed figures of merit that approximate it in one way or another.

In our experiments, we use only tag sequences for parsing. More accurate probability estimates should be attainable using lexical information.

2 Figures of Merit

2.1 Normalized β

It seems reasonable to base a figure of merit on the inside probability of the constituent. Inside probability is defined as the probability of the words or tags in the constituent given that the constituent is dominated by a particular nonterminal symbol. This seems to be a reasonable basis for comparing constituent probabilities, and has the additional advantage that it is easy to compute during chart parsing.

The inside probability of the constituent $N^i_{j,k}$ is given as:

$$\beta(N^i_{j,k}) \equiv p(t_{j,k}|N^i)$$

In terms of our earlier discussion, our "ideal" figure of merit can be rewritten as:

$$p(N^i_{j,k}|t_{0,n}) \quad = \quad \frac{\alpha(N^i_{j,k})\beta(N^i_{j,k})}{p(t_{0,n})}$$

where $\alpha(N^i_{j,k})$ and $p(t_{0,n})$ represent the influence of the surrounding words. Thus using β assumes that α and $p(t_{0,n})$ can be ignored. (The α term is explained in the next section.)

One side effect from omitting these terms is that inside probability alone tends to prefer shorter constituents to longer ones, as the inside probability of a longer constituent involves the product of more probabilities, and so some technique is used to normalize this quantity for use as a figure of merit. One approach is to take the geometric mean of the inside probability, to obtain a "per-word" inside probability. (In the "ideal" model, the $p(t_{0,n})$ term acts as a normalizing factor.)

The per-word inside probability of the constituent $N^i_{j,k}$ is calculated as

$$\sqrt[k-j]{\beta(N^i_{j,k})}$$

We will refer to this figure as **normalized** β.

2.2 Normalized $\alpha_L\beta$

Outside probability α of a constituent $N^i_{j,k}$ is defined as the probability of that constituent and the rest of the words in the sentence (or rest of the tags in the tag sequence, in our case), $p(t_{0,j}, N^i_{j,k}, t_{k,n})$. In some of our figures of merit, we used a closely related quantity, $p(N^i_{j,k}, t_{0,j})$, which we will call the left outside probability, or α_L. This term can be calculated recursively as follows:

$$\alpha_L(N^i_{j,k}) \quad = \quad \sum_{e \in \mathcal{E}^i_{j,k}} \alpha_L(N^{\mathrm{lhs}(e)}_{\mathrm{start}(e),\mathrm{end}(e)})p(\mathrm{rule}(e))\beta(N^q_{r,s})$$

where $\mathcal{E}^i_{j,k}$ is the set of all completed edges in which $N^i_{j,k}$ appears and r and s are defined by $N^q_{r,s} \in \mathrm{rhs}(e)$, $N^q_{r,s} \neq N^i_{j,k}$, and $r \leq j$. A method for calculating α_L can be derived from the calculations given in [7].

A simple extension to the normalized β model allows us to estimate the per-word probability of all tags in the sentence through the end of the constituent under consideration. This allows us to take advantage of information already obtained in a left-right parse. We calculate this quantity as follows:

$$\sqrt[k]{\alpha_L(N^i_{j,k})\beta(N^i_{j,k})}$$

We are again taking the geometric mean to compensate for the $\alpha\beta$ quantity's preference for shorter constituents, as explained in the previous section.

We refer to this figure of merit as **normalized** $\alpha_L\beta$.

2.3 Trigram estimate

We can rewrite the "ideal" figure of merit as follows:

$$\begin{aligned} p(N^i_{j,k}|t_{0,n}) &= \frac{p(N^i_{j,k}, t_{0,n})}{p(t_{0,n})} \\ &= \frac{p(t_{0,j}, t_{k,n})p(N^i_{j,k}|t_{0,j}, t_{k,n})p(t_{j,k}|N^i_{j,k}, t_{0,j}, t_{k,n})}{p(t_{0,j}, t_{k,n})p(t_{j,k}|t_{0,j}, t_{k,n})} \\ &\approx \frac{p(N^i_{j,k}|t_{0,j}, t_{k,n})p(t_{j,k}|N^i_{j,k})}{p(t_{j,k}|t_{0,j}, t_{k,n})} \end{aligned}$$

applying the usual independence assumption that given a nonterminal, the tag sequence it generates depends only on that nonterminal.

For this estimate, we make some additional independence assumptions. We assume that $p(N^i_{j,k}|t_{0,j}, t_{k,n}) \approx p(N^i_{j,k})$, that is, that the probability of a nonterminal is independent of the tags before and after it in the sentence. We also use a trigram model for the tags themselves, giving $p(t_{j,k}|t_{0,j}, t_{k,n}) \approx p(t_{j,k}|t_{j-2,j})$. Then we have:

$$p(N^i_{j,k}|t_{0,n}) \approx \frac{p(N^i)\beta(N^i_{j,k})}{p(t_{j,k}|t_{j-2,j})}$$

We can calculate $\beta(N^i_{j,k})$ as usual. The $p(N^i)$ term is estimated from our PCFG as the sum of the counts for all rules having N^i as their left-hand side, divided by the sum of the counts for all rules. The $p(t_{j,k}|t_{j-2,j})$ term is just the probability of the tag sequence $t_j \dots t_{k-1}$ according to a trigram (actually, tritag) model.[1] We refer to this model as the **trigram estimate**.

2.4 Prefix estimate

Another estimate we used took advantage of statistics on the first $j-1$ tags of the sentence as well, to consider the probability of the constituent in the context of the preceding tags.

$$
\begin{aligned}
p(N^i_{j,k}|t_{0,n}) &= \frac{p(N^i_{j,k}, t_{0,n})}{p(t_{0,n})} \\
&= \frac{p(t_{k,n})p(N^i_{j,k}, t_{0,j}|t_{k,n})p(t_{j,k}|N^i_{j,k}, t_{0,j}, t_{k,n})}{p(t_{k,n})p(t_{0,k}|t_{k,n})} \\
&= \frac{p(N^i_{j,k}, t_{0,j}|t_{k,n})p(t_{j,k}|N^i_{j,k}, t_{0,j}, t_{k,n})}{p(t_{0,k}|t_{k,n})}
\end{aligned}
$$

We again make the independence assumption that $p(t_{j,k}|N^i_{j,k}, t_{0,j}, t_{k,n}) \approx \beta(N^i_{j,k})$. Additionally, we assume that $p(N^i_{j,k}, t_{0,j})$ and $p(t_{0,k})$ are independent of $p(t_{k,n})$, giving

$$
p(N^i_{j,k}|t_{0,n}) \approx \frac{p(N^i_{j,k}, t_{0,j})\beta(N^i_{j,k})}{p(t_{0,k})}
$$

The denominator, $p(t_{0,k})$, is once again calculated from a tritag model. The $p(N^i_{j,k}, t_{0,j})$ term is computed as in the normalized $\alpha_L\beta$ model.

We will refer to this figure as the **prefix estimate**.

3 The Experiment

We used as our grammar a probabilistic context-free grammar learned from the Brown corpus (see [6], [2], [3], [4]). We parsed 500 sentences of length 3

[1] Our results show that the $p(N^i)$ term can be omitted without much effect.

to 30 (including punctuation) from the Penn Treebank Wall Street Journal corpus using a best-first parsing method and each of the following estimates for $p(N^i_{j,k}|t_{0,n})$ as the figure of merit:

1. normalized β
2. normalized $\alpha_L\beta$
3. trigram estimate
4. prefix estimate

Our tritag probabilities for the trigram and prefix estimates, as well as the probability $p(N^i)$ in the trigram estimate, were determined from the same training data from which our grammar was learned initially.

For each figure of merit, we compared the performance of best-first parsing using that figure of merit to exhaustive parsing. By exhaustive parsing, we mean continuing to parse until there are no more constituents available to be added to the chart. We parse exhaustively to determine the total probability of a sentence, that is, the sum of the probabilities of all parses found for that sentence.

We then computed several quantities for best-first parsing with each figure of merit at the point where the best-first parsing method has found parses contributing at least 95% of the probability mass of the sentence.

4 Results

The chart below presents the following measures for each figure of merit:

1. %E: The percentage of edges, or rule expansions, in the exhaustive parse that have been used by the best-first parse to get 95% of the probability mass. Edge creation is generally considered the best measure of CFG parser effort.
2. %non-0 E: The percentage of nonzero-length edges used by the best-first parse to get 95%. Zero-length edges are required by our parser as a book-keeping measure, and as such are virtually un-elimitable. We anticipated that removing them from consideration would highlight the "true" differences in the figures of merit.

3. %popped: The percentage of edges that were added to the keylist (i.e., the stack in a standard chart parse), that were added to the chart in order to get 95%

Figure of Merit	%E	%non-0 E	%popped
normalized β	34.7	31.6	61.5
normalized $\alpha_L\beta$	39.7	36.4	57.3
trigram estimate	25.2	21.7	44.3
prefix estimate	21.8	17.4	38.3

We gathered statistics for each sentence length from 3 to 30. Sentence length was limited to a maximum of 30 because of the huge number of edges that are generated in doing a full parse of long sentences; using this grammar, sentences in this length range have produced up to 130,000 edges. Figure 1 shows a graph of %non-0 E, that is, the percent of nonzero-length edges needed to get 95% of the probability mass, for each sentence length.

5 Previous work

Bobrow[1] and Chitrao and Grishman[5] introduced statistical agenda-based parsing techniques. Chitrao and Grishman implemented a best-first probabilistic parser and noted the parser's tendency to prefer shorter constituents. They proposed a heuristic solution of penalizing shorter constituents by a fixed amount per word.

Miller and Fox[11] compare the performance of parsers using three different types of grammars, and show that a probabilistic context-free grammar using inside probability (unnormalized) as a figure of merit outperforms both a context-free grammar and a context-dependent grammar.

Kochman and Kupin[8] propose a figure of merit closely related to our prefix estimate. They do not actually incorporate this figure into a best-first parser.

Magerman and Marcus [9] use the geometric mean to compute a figure of merit that is independent of constituent length. Magerman and Weir[10] use a similar model with a different parsing algorithm.

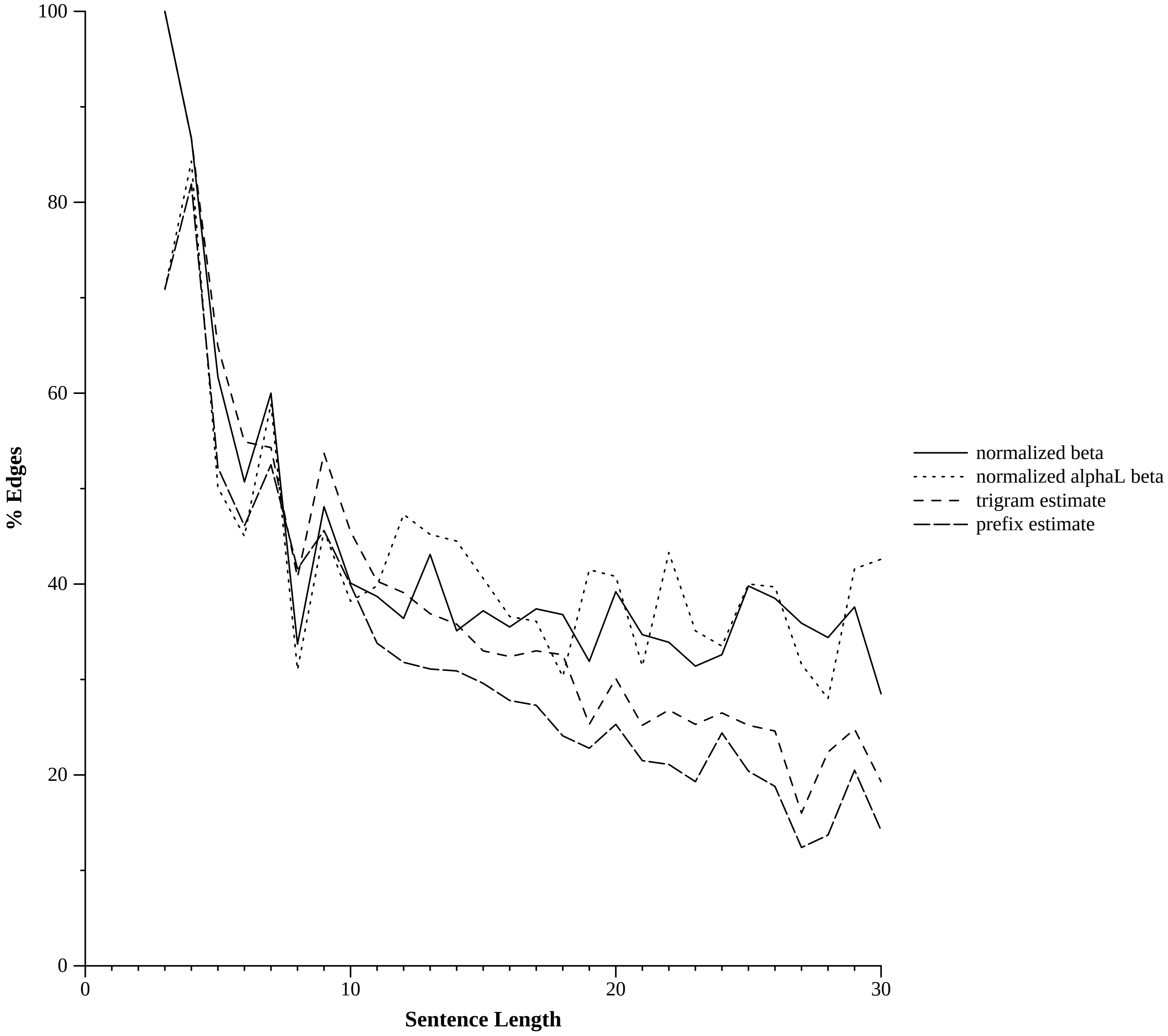

Figure 1: Nonzero-length edges for 95% of Probability Mass

6 Conclusions

The best performance in terms of edge counts of the figures we tested was the model which used the most information available from the sentence, the prefix model. However, so far, the additional running time needed for the computation of α_L terms has exceeded the time saved by processing fewer edges.

While chart parsing and calculations of β can be done in $O(n^3)$ time, we have been unable to find an algorithm to compute the α_L terms faster than $O(n^5)$. When a constituent is removed from the keylist, it only affects the β values of its ancestors in the parse trees; however, α_L values are propagated to all of the constituent's siblings to the right and all of its descendants. Recomputing the α_L terms when a constituent is removed from the keylist can be done in $O(n^3)$ time, and since there are $O(n^2)$ possible constituents, the total time needed to compute the α_L terms in this manner is $O(n^5)$.

The best performer in running time was the parser using the trigram estimate as a figure of merit. This figure has the additional advantage that it can be easily incorporated into existing best-first parsers using a figure of merit based on inside probability.

It is also interesting to note that while the models using figures of merit normalized by the geometric mean performed similarly to the other models on shorter sentences, the superior performance of the other models becomes more pronounced as sentence length increases. From Figure 1, we can see that the models using the geometric mean appear to level off with respect to an exhaustive parse when used to parse sentences of length greater than about 15. The other two estimates seem to continue improving with greater sentence length. In fact, the measurements presented here almost certainly underestimate the true benefits of the better models. We restricted sentence length to a maximum of 30 words, in order to keep the number of edges in the exhaustive parse to a practical size; however, since the percentage of edges needed by the best-first parse decreases with increasing sentence length, we assume that the improvement would be even more dramatic for sentences longer than 30 words.

References

[1] Bobrow, Robert J. (1990). Statistical Agenda Parsing. *DARPA Speech and Language Workshop*, 222-224.

[2] Carroll, Glenn and Eugene Charniak (1992). Learning probabilistic dependency grammars from labeled text. *Working Notes, Fall Symposium Series, AAAI*, 25-32.

[3] Carroll, Glenn and Eugene Charniak (1992). Two Experiments on Learning Probabilistic Dependency Grammars from Corpora. *Workshop Notes, Statistically-Based NLP Techniques, AAAI*, 1-13.

[4] Charniak, Eugene and Glenn Carroll (1994). Context-sensitive statistics for improved grammatical language models. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 728-733.

[5] Chitrao, Mahesh V. and Ralph Grishman (1990). Statistical Parsing of Messages. *DARPA Speech and Language Workshop*, 263-266.

[6] Francis, W. Nelson, and Henry Kučera (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

[7] Jelinek, Frederick and John D. Lafferty (1991). Computation of the Probability of Initial Substring Generation by Stochastic Context-Free Grammars. *Computational Linguistics, 17*, 315-323.

[8] Kochman, Fred and Joseph Kupin (1991). Calculating the Probability of a Partial Parse of a Sentence. *DARPA Speech and Language Workshop*, 237-240.

[9] Magerman, David M. and Mitchell P. Marcus (1991). Parsing the Voyager Domain Using Pearl. *DARPA Speech and Language Workshop*, 231-236.

[10] Magerman, David M. and Carl Weir (1992). Efficiency, Robustness and Accuracy in Picky Chart Parsing. *Proceedings of the 30th ACL Conference*, 40-47.

[11] Miller, Scott and Heidi Fox (1994). Automatic Grammar Acquisition. *Proceedings of the Human Language Technology Workshop*, 268-271.