Using the Penn Treebank to Evaluate Non-Treebank Parsers

Eric K. Ringger, Robert C. Moore,
Lucy Vanderwende, Hisami Suzuki
Microsoft Research
One Microsoft Way
Redmond, Washington 98027, USA
{ringger, bobmoore, lucyv, hisamis}@microsoft.com

Eugene Charniak
Computer Science
Box 1910
Brown University
Providence, RI 02912, USA
ec@cs.brown.edu

Abstract

This paper describes a method for conducting evaluations of Treebank and non-Treebank parsers alike against the English language U. Penn Treebank (Marcus et al., 1993) using a metric that focuses on the accuracy of relatively non-controversial aspects of parse structure. Our conjecture is that if we focus on maximal projections of heads (MPH), we are likely to find much broader agreement than if we try to evaluate based on order of attachment. We hope that this method may find wider acceptance and be useful in establishing a generally applicable framework for evaluation in natural language parsing. We employ this method in an evaluation of NLPWin (Heidorn, 2000), a parser developed at Microsoft Research without reference to the Penn Treebank, and, for comparison, the well-known statistical Treebank parser of Charniak (2000).

1. Introduction

In recent years, the literature includes compelling evaluation results for several natural language parsers that train and test on the U. Penn Treebank (Marcus et al., 1993), including the work of Magerman (1995), Collins (1999), Collins & Duffy (2002), Charniak (2000), and Klein & Manning (2003). Since the publication of the Treebank, published evaluations of non-Treebank parsers (i.e., parsers not designed with the conventions of the English language U. Penn Treebank in mind) are rarely encountered; recent exceptions include an evaluation of the Xerox LFG f-structure parser (Riezler, 2002) and the dependency parser of Lin (1995, 1998). When such evaluations do appear, the community does not regularly cite them and appears to be at a loss as to how to compare these results with the results published for the well-known Treebank parsers. We introduce a method for conducting evaluations of both Treebank and non-Treebank parsers using a metric that focuses on the accuracy of relatively non-controversial aspects of parse structure.

1.1. Data

We are using the English language University of Pennsylvania Treebank v.3 from the LDC. The Treebank contains parse trees annotated according to the Treebank annotation guidelines (Bies et al., 1995). Included in the Treebank are two datasets of interest to us: the Wall Street Journal (WSJ) and the Brown Corpus. The former has been divided by the parsing community into a standard training set (sections 02-21), a development test set (section 24), a blind test set (section 23), and some remainder sections. Charniak (2000), Collins (1997), and others use this standard division. To date, there is no standard division of the Brown Corpus Treebank, so we have provided one for this evaluation. We will elaborate on the division of the Brown Corpus Treebank after discussing the metric.

1.2. Non-Treebank Parsers

Natural language parsers not explicitly designed or trained to follow the conventions of the Penn Treebank may differ from the Treebank in any number of ways. Differences such as tokenization, part-of-speech labels, granularity of non-terminal constituents, and non-terminal constituent labels can usually be handled by ignoring labels and resorting to the “crossing brackets” metric, one of the Parseval metrics (Black, 1991). However, irreconcilable differences often remain, particularly with respect to whether left or right modifiers should attach to heads first. In situations where a parser has been designed to make attachments in an order different from that of the Treebank, large numbers of apparent crossing-bracket errors can result.

Consider an example from NLPWin, a parser developed at Microsoft Research (Heidorn, 2000) and the Treebank. Figure 1 is a tree produced by NLPWin, where each non-terminal represents a maximal projection of a single head, and each non-punctuation terminal is the immediate daughter of a non-terminal for which it is the head. Figure 2 contains the corresponding tree for the same sentence from the Treebank. There are several differences in tokenization, part-of speech tags, non-terminal constituents, and non-terminal labels.

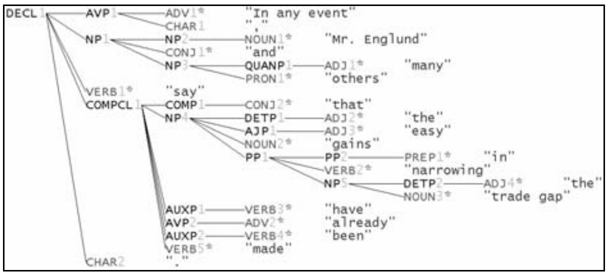
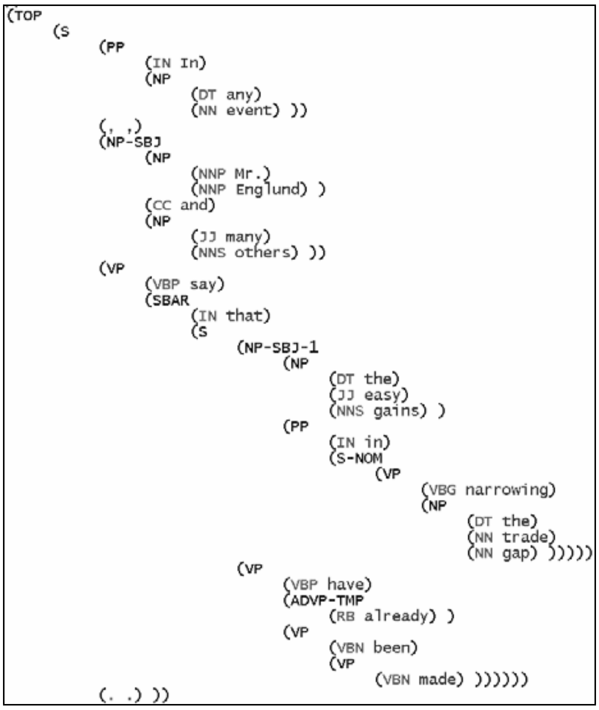


Figure 1: An NLPWin tree for a sentence
form the Wall Street Journal portion of the Treebank.



```
(TOP
    (S
        (PP
            (IN In)
            (NP
                (DT any)
                (NN event) ))
        (, ,)
        (NP-SBJ
            (NP
                (NNP Mr.)
                (NNP Englund) )
            (CC and)
            (NP
                (JJ many)
                (NNS others) ))
        (VP
            (VBP say)
            (SBAR
                (IN that)
                (S
                    (NP-SBJ-1
                        (NP
                            (DT the)
                            (JJ easy)
                            (NNS gains) )
                        (PP
                            (IN in)
                            (S-NOM
                                (VP
                                    (VBG narrowing)
                                    (NP
                                        (DT the)
                                        (NN trade)
                                        (NN gap) )))))
                    (VP
                        (VBP have)
                        (ADVP-TMP
                            (RB already) )
                        (VP
                            (VBN been)
                            (VP
                                (VBN made) ))))))
        (. .) ))
```

Figure 2: A bracketed tree (minus empty nodes) from the Wall Street Journal portion of the Treebank.

1.3. Maximal Projections of Heads

Faced with this lack of comparability between non-Treebank parses and Treebank parses, how can we proceed? Our conjecture is that if we focus on maximal projections of heads (MPH), we are likely to find much broader agreement than if we try to evaluate based on order of attachment or on the granularity of intermediate projections. While there are some theoretical differences that are not normalized by looking only at MPH brackets (e.g., whether "small clauses" exist or not), we proceed under the assumption that if only MPH brackets are evaluated, the remaining systematic differences are few enough to be dealt with on a case-by-case basis. In particular, to address tokenization issues, we added a post-process to NLPWin to tokenize in the manner of the Treebank, and we exclude part-of-speech tags and constituent labels from our evaluation. Differences with regard to the granularity of constituents, as well as order of attachment issues, are handled by focusing on maximal projections of heads.

A challenge posed by focusing on maximal projections of heads is the absence of annotations of heads and head inheritance in the Treebank itself.[1] To rectify this, we will employ a set of head-labeling rules and compute maximal projections of heads for the reference or "gold" Treebank trees. We can then have a comparable reference set containing only maximal projections against which to compute bracket precision and recall (Black, 1991), as well as crossing brackets. We elaborate on this procedure in the next section.

2. The Metrics and Evaluation Procedure

To compute MPH trees from the Treebank, we use a three-step process. First, we employ Charniak's head-labeling rules. There will inevitably be discrepancies between the resulting heads and the heads chosen by the parser to be evaluated. In general, one head assignment or the other produces strictly fewer MPH brackets; therefore, we always resolve such disagreements in favor of the head choice leading to fewer brackets, because the more detailed bracketing is not necessarily recoverable from the other source. We construct a small set of additional rules capturing the regularities in head reassignment. After heads are assigned, we compute the MPH brackets.

When evaluating a Treebank-style parser, we employ that parser's choice of heads, or if the parser does not produce head annotations, we can employ Charniak's rules. To facilitate a comparison with a non-Treebank parser, we then complete the annotation by applying the set of additional head reassignment rules constructed above.

It is worth noting that in computing maximal projections of heads for a given tree, no brackets are added; we keep only the brackets that enclose maximal projections of the specified heads. In fact, when changing head choices, we are simply focusing on a set of brackets that allows for a more direct comparison between a Treebank-style parser and a non-Treebank parser.[2]

After computation of MPH, to avoid giving credit for trivially recoverable brackets, we ignore brackets containing only a single terminal (as is standard) as well as the top level brackets around the entire sentence. In addition, in order to avoid free credit for re-labeling non-terminals, we flatten all non-branching non-terminals. It is standard to eliminate punctuation from the hypothesis and reference trees just before evaluating. Consequently, we ignore terminals with punctuation tags by employing the appropriate options in the EvalB evaluation tool (Sekine & Collins, 1997). With these refinements, our metrics are, therefore, precision and recall of unlabeled brackets enclosing maximal projections of heads (MPH), defined as follows:

Unlabeled MPH Precision

$$= \frac{\text{\# of correct constituents in proposed MPH tree}}{\text{\# of constituents in proposed MPH tree}}$$

Unlabeled MPH Recall

$$= \frac{\text{\# of correct constituents in proposed MPH tree}}{\text{\# of constituents in Trebank MPH tree}}$$

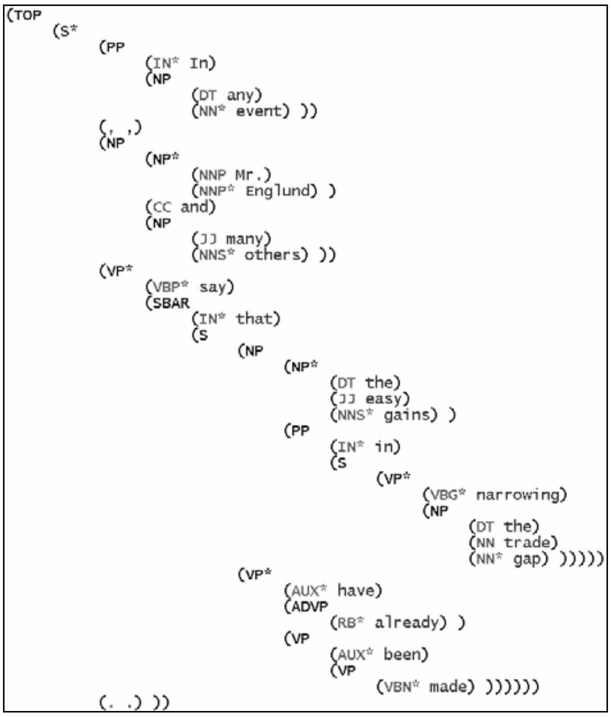
[1] We observe that if we leave the Treebank trees untouched and compute only the precision of MPH trees against the untouched Treebank, we introduce a scoring bias in favor of high PP attachment; i.e., we would give credit to false high attachments.

[2] The set of additional head reassignment rules for one non-Treebank parser is likely to be different than the rule set for another such parser. A comparison between two non-Treebank parsers would thus require constructing a set of head reassignment rules capturing the common-ground among the Charniak-head-labeled Treebank and the two parsers.

We recommend a comparison of our approach with the proposals of Gaizauskas et al. (1998a, 1998b); the principal contrast lies in our employment of head labeling rules and MPH.

3. Comparison System

As a point of comparison with NLPWin and to provide a connection with past Treebank evaluations, we use Eugene Charniak's state of the art statistical Treebank-style parser (Charniak, 2000). Charniak's head-driven parser produces trees with heads. Figure 3 depicts the tree from Charniak's parser for the sentence used in the example trees above. Asterisks (*) mark the heads or daughters through which heads are inherited. For the sake of a direct comparison with NLPWin, we apply the set of head reassignment rules (found as the common-ground between NLPWin and the Charniak-head-labeled Treebank) and compute the MPH trees using the process defined in section 2 to arrive at the tree in Figure 4. The primary changes are in the structure of the PP "in narrowing the trade gap" and in the head for the final S constituent "… have already been made".



```
(TOP
    (S*
        (PP
            (IN* In)
            (NP
                (DT any)
                (NN* event) ))
        (, ,)
        (NP
            (NP*
                (NNP Mr.)
                (NNP* Englund) )
            (CC and)
            (NP
                (JJ many)
                (NNS* others) ))
        (VP*
            (VBP* say)
            (SBAR
                (IN* that)
                (S
                    (NP
                        (NP*
                            (DT the)
                            (JJ easy)
                            (NNS* gains) )
                        (PP
                            (IN* in)
                            (S
                                (VP*
                                    (VBG* narrowing)
                                    (NP
                                        (DT the)
                                        (NN trade)
                                        (NN* gap) )))))
                    (VP*
                        (AUX* have)
                        (ADVP
                            (RB* already) )
                        (VP
                            (AUX* been)
                            (VP
                                (VBN* made) ))))))
        (. .) ))
```

Figure 3: A parse tree from Charniak's parser; heads and head inheritance marked with asterisks.

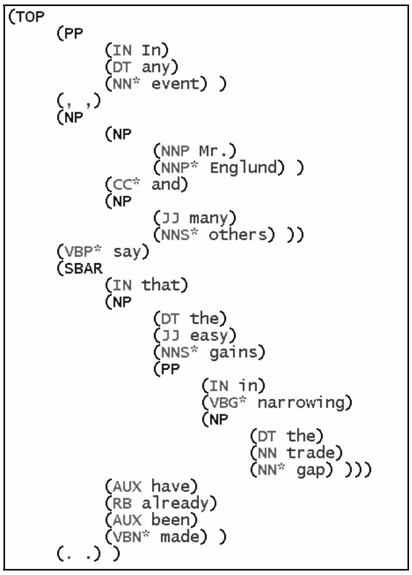
4. Evaluation Results

4.1 WSJ Treebank

We evaluated NLPWin MPH trees and MPH trees from Charniak's parser side-by-side. For each result, we include the statistics listed in the first column of Table 1, most of them computed by EvalB in the standard ways. Columns 2-3 contain the primary results on the WSJ development test set (section 24), and columns 4-5 are the blind test set (section 23) results. Every sentence is included in these results without regard to sentence length.

The number of NLPWin HYP brackets and the number of Charniak HYP brackets are quite close to one another and to the number of REF brackets, which is one source of our confidence that we are indeed evaluating the two different systems in a comparable way against comparable reference brackets.

In brief, the results are as follows: on the blind test section (section 23) of the Wall Street Journal portion of the Treebank, NLPWin achieves precision (P) of 74.32 and recall (R) of 74.25. For Charniak: P=86.18, R=85.72.



```
(TOP
    (PP
        (IN In)
        (DT any)
        (NN* event) )
    (, ,)
    (NP
        (NP
            (NNP Mr.)
            (NNP* Englund) )
        (CC* and)
        (NP
            (JJ many)
            (NNS* others) ))
    (VBP* say)
    (SBAR
        (IN that)
        (NP
            (DT the)
            (JJ easy)
            (NNS* gains)
            (PP
                (IN in)
                (VBG* narrowing)
                (NP
                    (DT the)
                    (NN trade)
                    (NN* gap) )))
        (AUX have)
        (RB already)
        (AUX been)
        (VBN* made) )
    (. .) )
```

Figure 4: The tree of Figure 3 after selecting heads for agreement with NLPWin and as MPH.

4.2 Restricted Brown Corpus Treebank

The standard WSJ sections provide a benchmark, but since Charniak's parser was trained on WSJ material and NLPWin was not, we sought a data set that was equally novel to both parsers. We are using part of the Brown Corpus section of the Treebank (hereafter "the Brown Corpus Treebank" or BCTB) to play this role.

A small amount of the original Brown Corpus (Francis & Kucera, 1967) material had been used in developing NLPWin, so we removed that subset from the BCTB, calling the remainder the "Restricted BCTB". We randomized the Restricted BCTB[3], divided it into sections of 2000 sentences, and named a development test section (sec. BR0) and a blind test section (sec. BR1). The data that had been used in development of NLPWin was provided to Charniak, and he retrained his parser incorporating the used data.

The results of evaluating both NLPWin and Charniak's parser on BR0 are in columns 6-7, and results on BR1 are in columns 8-9 of Table 1. As before, every sentence is included in these results without regard to sentence length.

[3] The Randomized, Restricted, Brown Corpus (RRBC) Treebank is the data set we are truly interested in for our evaluation. This set consists of 22,021 sentences from the 24,243 sentences in the Brown Corpus Treebank.

	WSJ 24		WSJ 23		BR 0		BR 1	
	NLPWin	Charniak	NLPWin	Charniak	NLPWin	Charniak	NLPWin	Charniak
Number of sentences	1346	1346	2416	2416	2000	2000	2000	2000
Number of reference brackets in the Treebank MPH trees	10494	10496	17686	17688	11864	11868	11645	11648
Number of hypothesis brackets in the MPH trees proposed by the respective parser	10407	10380	17669	17594	12043	11819	11816	11651
Number of hypothesis MPH brackets that match reference MPH brackets	7768	8903	13131	15163	8489	9415	8217	9054
Unlabeled bracket recall	**74.02**	**84.82**	**74.25**	**85.72**	**71.55**	**79.33**	**70.56**	**77.73**
Unlabeled bracket precision	**74.64**	**85.77**	**74.32**	**86.18**	**70.49**	**79.66**	**69.54**	**77.71**
Number of sentences with complete match of hypothesis and reference	23.92	41.23	27.11	45.16	32.95	45.75	32.8	44.3
Average number of crossing brackets / sent.	0.72	0.32	0.7	0.29	0.67	0.4	0.67	0.47
Number of sentences with no crossing brack.	69.24	80.91	68.34	82.41	73	80.55	73.2	79.05
Number of sentences with 2 or fewer crossing brackets	90.04	97.03	90.44	97.14	91.3	95.65	90.85	94.5

Table 1: Summary of evaluation results for NLPWin and Charniak's parser on the following data sets: development test set WSJ24, blind test set WSJ23, development test set BR0, and blind test set BR1.

Briefly, for the blind test section (BR1), the two systems scored as follows: for NLPWin: P=69.54, R=70.56; for Charniak: P=77.71, R=77.73, a narrower performance gap than we observed on the WSJ. Unlike the WSJ results, the number of NLPWin HYP brackets in column 6 is somewhat larger than both the number of REF brackets in that column and the number of HYP brackets in column 7; we do not observe a regular trend responsible for this difference.

5. Conclusions

We have introduced a method for conducting evaluations of both Treebank and non-Treebank parsers using a metric that focuses on the accuracy of relatively non-controversial aspects of parse structure, namely unlabeled precision and recall of maximal projections of heads. We hope that this method may find wider acceptance and be useful in establishing a generally applicable framework for evaluation in natural language parsing. Note that since a simple dependency tree (with ordered dependents) is isomorphic to an MPH tree (with heads labeled), this work also opens the door to evaluating dependency parsers (such as those described in the work of Riezler and of Lin cited above) against the Treebank.

Acknowledgments

Thanks go to David Parkinson for scrutiny of many trees for verification of the head selection process, to Karen Jensen, Lee Schwartz, and Rich Campbell for the NLPWin English grammar and related rules. Thanks also go to the anonymous reviewers, one of whom alerted us to the closely related work of Gaizauskas et al.

References

Bies, A., M. Ferguson, K. Katz, R. MacIntyre, V. Tredinnick, G. Kim, M.A. Marcinkiewicz, B. Schasberger. (1995) "Bracketing Guidelines for Treebank II Style Penn Treebank Project". TR, University of Pennsylvania.

Black, E et al. (1991) "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars". In Proceedings of the February 1991 DARPA Speech and Natural Language Workshop. pp. 306-311.

Charniak, E. (2000) "A Maximum-Entropy-Inspired Parser". In Proceedings of NAACL-2000. pp. 132–139.

Collins, M. (1999) "Head-Driven Statistical Models for Natural Language Parsing". Ph.D. thesis, U. Pennsylvania.

Collins, M. & N. Duffy. (2002) "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron". In Proceedings of ACL 2002.

Gaizauskas, R., M. Hepple, & C. Huyck. (1998a) "A Scheme for Comparative Evaluation of Diverse Parsing Systems". In Proceedings of 1st International Conference on Language Resources and Evaluation (LREC'98), pp. 143-149.

Gaizauskas, R., M. Hepple, & C. Huyck. (1998b) "Modifying Existing Annotated Corpora for General Comparative Evaluation of Parsing". In Proceedings of the LREC'98 Workshop on Evaluation of Parsing Systems.

Heidorn, G. (2000). "Intelligent writing assistance". In Dale et al. *Handbook of Natural Language Processing*, Marcel Dekker.

Kucera, H. & W.N. Francis. (1967) *Computational Analysis of Present-Day American English.* Providence: Brown University Press.

Lin, D. (1995) "A dependency-based method for evaluating broad-coverage parsers". In Proceedings of IJCAI-95, pp.1420-1425.

Lin, D. (1998) "Dependency-based Evaluation of MINIPAR". In Workshop on the Evaluation of Parsing Systems.

Magerman, D. (1995) "Statistical decision-tree models for parsing". In ACL 33, pp. 276–283.

Marcus, M., B. Santorini, M.A. Marcinkiewicz. (1993) "Building a Large Annotated Corpus of English: The Penn Treebank". *Computational Linguistics* 19 (2), 313-330.

Riezler, S., T. King, R. Kaplan, R. Crouch, J. Maxwell III, M. Johnson. (2002) "Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques". In Proceedings of ACL 2002. pp. 271-278.

Sekine, S. & M. Collins. (1997) EvalB: a bracket scoring program. URL: http://www.cs.nyu.edu/cs/projects/proteus/evalb/