Corpus Variation and Parser Performance

Daniel Gildea
University of California, Berkeley, and
International Computer Science Institute
gildea@cs.berkeley.edu

Abstract

Most work in statistical parsing has focused on a single corpus: the Wall Street Journal portion of the Penn Treebank. While this has allowed for quantitative comparison of parsing techniques, it has left open the question of how other types of text might affect parser performance, and how portable parsing models are across corpora. We examine these questions by comparing results for the Brown and WSJ corpora, and also consider which parts of the parser's probability model are particularly tuned to the corpus on which it was trained. This leads us to a technique for pruning parameters to reduce the size of the parsing model.

1 Introduction

The past several years have seen great progress in the field of natural language parsing, through the use of statistical methods trained using large corpora of hand-parsed training data. The techniques of Charniak (1997), Collins (1997), and Ratnaparkhi (1997) achieved roughly comparable results using the same sets of training and test data. In each case, the corpus used was the Penn Treebank's hand-annotated parses of Wall Street Journal articles. Relatively few quantitative parsing results have been reported on other corpora (though see Stolcke et al. (1996) for results on Switchboard, as well as Collins et al. (1999) for results on Czech and Hwa (1999) for bootstrapping from WSJ to ATIS). The inclusion of parses for the Brown corpus in the Penn Treebank allows us to compare parser performance across corpora. In this paper we examine the following questions:

- To what extent is the performance of statistical parsers on the WSJ task due to its relatively uniform style, and how might such parsers fare on the more varied Brown corpus?
- Can training data from one corpus be applied to parsing another?
- What aspects of the parser's probability model are particularly tuned to one corpus, and which are more general?

Our investigation of these questions leads us to a surprising result about parsing the WSJ corpus: over a third of the model's parameters can be eliminated with little impact on performance. Aside from cross-corpus considerations, this is an important finding if a lightweight parser is desired or memory usage is a consideration.

2 Previous Comparisons of Corpora

A great deal of work has been done outside of the parsing community analyzing the variations between corpora and different genres of text. Biber (1993) investigated variation in a number syntactic features over genres, or registers, of language. Of particular importance to statistical parsers is the investigation of frequencies for verb subcategorizations such as Roland and Jurafsky (1998). Roland et al. (2000) find that subcategorization frequencies for certain verbs vary significantly between the Wall Street Journal corpus and the mixed-genre Brown corpus, but that they vary less so between genre-balanced British and American corpora. Argument structure is essentially the task that automatic parsers attempt to solve, and the frequencies of various structures in training data are reflected in a statistical parser's probability model. The variation in verb argument structure found by previous research caused us to wonder to what extent a model trained on one corpus would be useful in parsing another. The probability models of modern parsers include not only the number and syntactic type of a word's arguments, but lexical information about their fillers. Although we are not aware of previous comparisons of the frequencies of argument fillers, we can only assume that they vary at least as much as the syntactic subcategorization frames.

3 The Parsing Model

We take as our baseline parser the statistical model of Model 1 of Collins (1997). The model is a history-based, generative model, in which the probability for a parse tree is found by expanding each node in the tree in turn into its child nodes, and multiplying the probabilities for each action in the derivation. It can

be thought of as a variety of lexicalized probabilistic context-free grammar, with the rule probabilities factored into three distributions. The first distribution gives probability of the syntactic category H of the head child of a parent node with category P, head word Hhw with the head tag (the part of speech tag of the head word) Hht:

$$P_h(H|P, Hht, Hhw)$$

The head word and head tag of the new node H are defined to be the same as those of its parent. The remaining two distributions generate the non-head children one after the other. A special #STOP# symbol is generated to terminate the sequence of children for a given parent. Each child is generated in two steps: first its syntactic category C and head tag Cht are chosen given the parent's and head child's features and a function Δ representing the distance from the head child:

$$P_c(C, Cht|P, H, Hht, Hhw, \Delta)$$

Then the new child's head word Chw is chosen:

$$P_{cw}(Chw|P, H, Hht, Hhw, \Delta, C, Cht)$$

For each of the three distributions, the empirical distribution of the training data is interpolated with less specific backoff distributions, as we will see in Section 5. Further details of the model, including the distance features used and special handling of punctuation, conjunctions, and base noun phrases, are described in Collins (1999).

The fundamental features of used in the probability distributions are the lexical heads and head tags of each constituent, the co-occurrences of parent nodes and their head children, and the co-occurrences of child nodes with their head siblings and parents. The probability models of Charniak (1997), Magerman (1995) and Ratnaparkhi (1997) differ in their details but are based on similar features. Models 2 and 3 of Collins (1997) add some slightly more elaborate features to the probability model, as do the additions of Charniak (2000) to the model of Charniak (1997).

Our implementation of Collins' Model 1 performs at 86% precision and recall of labeled parse constituents on the standard Wall Street Journal training and test sets. While this does not reflect the state-of-the-art performance on the WSJ task achieved by the more the complex models of Charniak (2000) and Collins (2000), we regard it as a reasonable baseline for the investigation of corpus effects on statistical parsing.

4 Parsing Results on the Brown Corpus

We conducted separate experiments using WSJ data, Brown data, and a combination of the two as training material. For the WSJ data, we observed the standard division into training (sections 2 through 21 of the treebank) and test (section 23) sets. For the Brown data, we reserved every tenth sentence in the corpus as test data, using the other nine for training. This may underestimate the difficulty of the Brown corpus by including sentences from the same documents in training and test sets. However, because of the variation within the Brown corpus, we felt that a single contiguous test section might not be representative. Only the subset of the Brown corpus available in the Treebank II bracketing format was used. This subset consists primarily of various fiction genres. Corpus sizes are shown in Table 1.

	Training Set		Test Set	
Corpus	Sentences	Words	Sentences	Words
WSJ	39,832	950,028	2245	48,665
Brown	21,818	413,198	2282	38,109

Table 1: Corpus sizes. Both test sets were restricted to sentences of 40 words or less. The Brown test set's average sentence was shorter despite the length restriction.

Training Data	Test Set	Recall	Prec.
WSJ	WSJ	86.1	86.6
WSJ	Brown	80.3	81.0
Brown	Brown	83.6	84.6
WSJ+Brown	Brown	83.9	84.8
WSJ+Brown	WSJ	86.3	86.9

Table 2: Parsing results by training and test corpus

Results for the Brown corpus, along with WSJ results for comparison, are shown in Table 2. The basic mismatch between the two corpora is shown in the significantly lower performance of the WSJ-trained model on Brown data than on WSJ data (rows 1 and 2). A model trained on Brown data only does significantly better, despite the smaller size of the training set. Combining the WSJ and Brown training data in one model improves performance further, but by less than 0.5% absolute. Similarly, adding the Brown data to the WSJ model increased performance on WSJ by less than 0.5%. Thus, even a large amount of additional data seems to have relatively little impact if it is not matched to the test material.

The more varied nature of the Brown corpus also seems to impact results, as all the results on Brown are lower than the WSJ result.

5 The Effect of Lexical Dependencies

The parsers cited above all use some variety of lexical dependency feature to capture statistics on the co-

occurrence of pairs of words being found in parent-child relations within the parse tree. These word pair relations, also called lexical bigrams (Collins, 1996), are reminiscent of dependency grammars such as Melčuk (1988) and the link grammar of Sleator and Temperley (1993). In Collins' Model 1, the word pair statistics occur in the distribution

$$P_{cw}(Chw|P, H, Hht, Hhw, \Delta, C, Cht)$$

where Hhw represent the head word of a parent node in the tree and Chw the head word of its (non-head) child. (The head word of a parent is the same as the head word of its head child.) Because this is the only part of the model that involves pairs of words, it is also where the bulk of the parameters are found. The large number of possible pairs of words in the vocabulary make the training data necessarily sparse. In order to avoid assigning zero probability to unseen events, it is necessary to smooth the training data. The Collins model uses linear interpolation to estimate probabilities from empirical distributions of varying specificities:

$$\begin{aligned} P_{cw}(Chw|P, H, Hht, Hhw, \Delta, C, Cht) = \\ \lambda_1 \tilde{P}(Chw|P, H, Hht, Hhw, \Delta, C, Cht) + \\ (1-\lambda_1)\Big(\lambda_2 \tilde{P}(Chw|P, H, Hht, \Delta, C, Cht) + \\ (1-\lambda_2)\tilde{P}(Chw|Cht)\Big) \end{aligned} \quad (1)$$

where $\tilde{P}$ represents the empirical distribution derived directly from the counts in the training data. The interpolation weights λ_1, λ_2 are chosen as a function of the number of examples seen for the conditioning events and the number of unique values seen for the predicted variable. Only the first distribution in this interpolation scheme involves pairs of words, and the third component is simply the probability of a word given its part of speech.

Because the word pair feature is the most specific in the model, it is likely to be the most corpus-specific. The vocabularies used in corpora vary, as do the word frequencies. It is reasonable to expect word co-occurrences to vary as well. In order to test this hypothesis, we removed the distribution $\tilde{P}(Chw|P, H, Hht, Hhw, C, Cht)$ from the parsing model entirely, relying on the interpolation of the two less specific distributions in the parser:

$$\begin{aligned} P_{cw2}(Chw|P, H, Hht, \Delta, C, Cht) = \\ \lambda_2 \tilde{P}(Chw|P, H, Hht, \Delta, C, Cht) + \\ (1-\lambda_2)\tilde{P}(Chw|Cht) \end{aligned} \quad (2)$$

We performed cross-corpus experiments as before to determine whether the simpler parsing model might be more robust to corpus effects. Results are shown in Table 3.

Perhaps the most striking result is just how little the elimination of lexical bigrams affects the baseline system: performance on the WSJ corpus decreases by less than 0.5% absolute. Moreover, the performance of a WSJ-trained system without lexical bigrams on Brown test data is identical to the WSJ-trained system with lexical bigrams. Lexical co-occurrence statistics seem to be of no benefit when attempting to generalize to a new corpus.

6 Pruning Parser Parameters

The relatively high performance of a parsing model with no lexical bigram statistics on the WSJ task led us to explore whether it might be possible to significantly reduce the size of the parsing model by selectively removing parameters without sacrificing performance. Such a technique reduces the parser's memory requirements as well as the overhead of loading and storing the model, which could be desirable for an application where limited computing resources are available.

Significant effort has gone into developing techniques for pruning statistical language models for speech recognition, and we borrow from this work, using the weighted difference technique of Seymore and Rosenfeld (1996). This technique applies to any statistical model which estimates probabilities by *backing off*, that is, using probabilities from a less specific distribution when no data are available are available for the full distribution, as the following equations show for the general case:

$$\begin{aligned} P(e|h) = & \quad P_1(e|h) & \text{if } e \notin \text{BO}(h) \\ = & \quad \alpha(h)P_2(e|h') & \text{if } e \in \text{BO}(h) \end{aligned}$$

Here e is the event to be predicted, h is the set of conditioning events or *history*, α is a backoff weight, and h' is the subset of conditioning events used for the less specific backoff distribution. BO is the backoff set of events for which no data are present in the specific distribution P_1. In the case of n-gram language modeling, e is the next word to be predicted, and the conditioning events are the $n-1$ preceding words. In our case the specific distribution P_1 of the backoff model is P_{cw} of equation 1, itself a linear interpolation of three empirical distributions from the training data. The less specific distribution P_2 of the backoff model is P_{cw2} of equation 2, an interpolation of two empirical distributions. The backoff weight α is simply $1-\lambda_1$ in our linear interpolation model. The Seymore/Rosenfeld pruning technique can be used to prune backoff probability models regardless of whether the backoff weights are derived from linear interpolation weights or discounting techniques such as Good-Turing. In order to ensure that the model's probabilities still sum to one, the backoff

Training Data	Test Set	w/ bigrams Recall	w/ bigrams Prec.	w/o bigrams Recall	w/o bigrams Prec.
WSJ	WSJ	86.1	86.6	85.6	86.2
WSJ	Brown	80.3	81.0	80.3	81.0
Brown	Brown	83.6	84.6	83.5	84.4
WSJ+Brown	Brown	83.9	84.8	83.4	84.3
WSJ+Brown	WSJ	86.3	86.9	85.7	86.4

Table 3: Parsing results by training and test corpus

weight α must be adjusted whenever a parameter is removed from the model. In the Seymore/Rosenfeld approach, parameters are pruned according to the following criterion:

$$N(e,h)(\log p(e|h) - \log p'(e|h')) \qquad (3)$$

where $p'(e|h')$ represents the new backed off probability estimate after removing $p(e|h)$ from the model and adjusting the backoff weight, and $N(e,h)$ is the count in the training data. This criterion aims to prune probabilities that are similar to their backoff estimates, and that are not frequently used. As shown by Stolcke (1998), this criterion is an approximation of the relative entropy between the original and pruned distributions, but does not take into account the effect of changing the backoff weight on other events' probabilities.

Adjusting the threshold θ below which parameters are pruned allows us to successively remove more and more parameters. Results for different values of θ are shown in Table 4.

The complete parsing model derived from the WSJ training set has 735,850 parameters in a total of nine distributions: three levels of backoff for each of the three distributions P_h, P_c and P_{cw}. The lexical bigrams are contained in the most specific distribution for P_{cw}. Removing all these parameters reduces the total model size by 43%. The results show a gradual degradation as more parameters are pruned.

The ten lexical bigrams with the highest scores for the pruning metric are shown in Table 5 for WSJ and Table 6. The pruning metric of equation 3 has been normalized by corpus size to allow comparison between WSJ and Brown. The only overlap between the two sets is for pairs of unknown word tokens. The WSJ bigrams are almost all specific to finance, are all word pairs that are likely to appear immediately adjacent to one another, and are all children of the base NP syntactic category. The Brown bigrams, which have lower correlation values by our metric, include verb/subject and preposition/object relations and seem more broadly applicable as a model of English. However, the pairs are not strongly related semantically, no doubt because the first term of the pruning criterion favors the most frequent words, such as forms of the verbs "be" and "have".

Child word *Chw*	*Head word* *Hhw*	*Parent* *P*	*Pruning* *Metric*
New	York	NPB	.0778
Stock	Exchange	NPB	.0336
$< unk >$	$< unk >$	NPB	.0313
vice	president	NPB	.0312
Wall	Street	NPB	.0291
San	Francisco	NPB	.0291
York	Stock	NPB	.0243
Mr.	$< unk >$	NPB	.0241
third	quarter	NPB	.0227
Dow	Jones	NPB	.0227

Table 5: Ten most significant lexical bigrams from WSJ, with parent category (other syntactic context variables not shown) and pruning metric
. NPB is Collins' "base NP" category.

Child word *Chw*	*Head word* *Hhw*	*Parent* *P*	*Pruning* *Metric*
It	was	S	.0174
it	was	S	.0169
$< unk >$	of	PP	.0156
$< unk >$	in	PP	.0097
course	Of	PP	.0090
been	had	VP	.0088
$< unk >$	$< unk >$	NPB	.0079
they	were	S	.0077
I	'm	S	.0073
time	at	PP	.0073

Table 6: Ten most significant lexical bigrams from Brown

7 Conclusion

Our results show strong corpus effects for statistical parsing models: a small amount of matched training data appears to be more useful than a large amount of unmatched data. The standard WSJ task seems to be simplified by its homogenous style. Adding training data from from an unmatched corpus doesn't hurt, but doesn't help a great deal either.

In particular, lexical bigram statistics appear to be corpus-specific, and our results show that they

Threshold θ	# parameters removed	% reduction model size	Recall	Prec.
0 (full model)	0	0	86.1	86.6
1	96K	13	86.0	86.4
2	166K	23	85.9	86.2
3	213K	29	85.7	86.2
∞	316K	43	85.6	86.2

Table 4: Parsing results with pruned probability models. The complete parsing model contains 736K parameters in nine distributions. Removing all lexical bigram parameters reducing the size of the model by 43%.

are of no use when attempting to generalize to new training data. In fact, they are of surprisingly little benefit even for matched training and test data — removing them from the model entirely reduces performance by less than 0.5% on the standard WSJ parsing task. Our selective pruning technique allows for a more fine grained tuning of parser model size, and would be particularly applicable to cases where large amounts of training data are available but memory usage is a consideration. In our implementation, pruning allowed models to run within 256MB that, unpruned, required larger machines.

The parsing models of Charniak (2000) and Collins (2000) add more complex features to the parsing model that we use as our baseline. An area for future work is investigation of the degree to which such features apply across corpora, or, on the other hand, further tune the parser to the peculiarities of the Wall Street Journal. Of particular interest are the automatic clusterings of lexical co-occurrences used in Charniak (1997) and Magerman (1995). Cross-corpus experiments could reveal whether these clusters uncover generally applicable semantic categories for the parser's use.

Acknowledgments This work was undertaken as part of the FrameNet project at ICSI, with funding from National Science Foundation grant ITR/HCI #0086132.

References

Douglas Biber. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–241, June.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *AAAI97*, Brown University, Providence, Rhode Island, August.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*, Seattle, Washington.

Michael Collins, Jan Hajic, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for czech. In *Proceedings of the 37th Annual Meeting of the ACL*, College Park, Maryland.

Michael Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the ACL.*

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the ACL.*

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing.* Ph.D. thesis, University of Pennsylvania, Philadelphia.

Michael Collins. 2000. Discriminative reranking for natural language parsing. In *Proceedings of the ICML.*

Rebecca Hwa. 1999. Supervised grammar induction using training data with limited constituent information. In *Proceedings of the 37th Annual Meeting of the ACL*, College Park, Maryland.

David Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the ACL.*

Ivan A. Meĺčuk. 1988. *Dependency Syntax: Theory and Practice.* State University of New York Press.

Adwait Ratnaparkhi. 1997. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing.*

Douglas Roland and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of COLING/ACL*, pages 1122–1128.

Douglas Roland, Daniel Jurafsky, Lise Menn, Susanne Gahl, Elizabeth Elder, and Chris Riddoch. 2000. Verb subcategorization frequency differences between business-news and balanced corpora: the role of verb sense. In *Proceedings of the Association for Computational Linguistics (ACL-2000) Workshop on Comparing Corpora.*

Kristie Seymore and Roni Rosenfeld. 1996. Scalable backoff language models. In *ICSLP-96*, volume 1, pages 232–235, Philadelphia.

Daniel Sleator and Davy Temperley. 1993. Parsing english with a link grammar. In *Third International Workshop on Parsing Technologies*, August.

A. Stolcke, C. Chelba, D. Engle, V. Jimenez, L. Mangu, H. Printz, E. Ristad, R. Rosenfeld, D. Wu, F. Jelinek, and S. Khudanpur. 1996. Dependency language modeling. Summer Workshop Final Report 24, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, April.

Andreas Stolcke. 1998. Entropy-based pruning of backoff language models. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274, Lansdowne, Va.