Learning Structured Classifiers for Statistical Dependency Parsing

Qin Iris Wang
Department of Computing Science
University of Alberta
Edmonton, Canada T6G 2E8
wqin@cs.ualberta.ca

Abstract

My research is focused on developing machine learning algorithms for inferring dependency parsers from language data. By investigating several approaches I have developed a unifying perspective that allows me to share advances between both probabilistic and non-probabilistic methods. First, I describe a generative technique that uses a strictly lexicalised parsing model, where all the parameters are based on words and do not use any part-of-speech (POS) tags nor grammatical categories. Then, I incorporate two ideas from probabilistic parsing—word similarity smoothing and local estimation—to improve the large margin approach. Finally, I present a simpler and more efficient approach to training dependency parsers by applying a boosting-like procedure to standard training methods.

1 Introduction

Over the past decade, there has been tremendous progress on learning parsing models from treebank data (Magerman, 1995; Collins, 1999; Charniak, 1997; Ratnaparkhi, 1999; Charniak, 2000; Wang et al., 2005; McDonald et al., 2005). Most of the early work in this area was based on postulating *generative* probability models of language that included parse structures (Magerman, 1995; Collins, 1997; Charniak, 1997). Learning in this context consisted of estimating the parameters of the model with simple likelihood based techniques, but incorporating various smoothing and back-off estimation tricks to cope with the sparse data problems (Collins, 1997; Bikel, 2004). Subsequent research began to focus more on *conditional* models of parse structure given the input sentence, which allowed discriminative training techniques such as maximum conditional likelihood (i.e. "maximum entropy") to be applied (Ratnaparkhi, 1999; Charniak, 2000). Currently, the work on conditional parsing models appears to have culminated in large margin training approaches (Taskar et al., 2004; McDonald et al., 2005), which demonstrates the state of the art performance in English dependency parsing.

Despite the realization that maximum margin training is closely related to maximum conditional likelihood for conditional models (McDonald et al., 2005), a sufficiently unified view has not yet been achieved that permits the easy exchange of improvements between the probabilistic and non-probabilistic approaches. For example, smoothing methods have played a central role in probabilistic approaches (Collins, 1997; Wang et al., 2005), and yet they are not being used in current large margin training algorithms. Another unexploited connection is that probabilistic approaches pay closer attention to the individual errors made by each component of a parse, whereas the training error minimized in the large margin approach—the "structured margin loss" (McDonald et al., 2005)—is a coarse measure that only assesses the total error of an entire parse rather than focusing on the error of any particular component. I have addressed both of these issues, as well as others in my work.

2 Dependency Parsing Model

Given a sentence $W = (w_1, ..., w_n)$, I consider the problem of computing an accurate directed depen-

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 5–8,
Rochester, April 2007.

dency tree, T, over W. Note that T consists of ordered pairs of words $(w_i \rightarrow w_j)$ in W such that each word appears in at least one pair and each word has in-degree at most one. Dependency trees are usually assumed to be projective (no crossing arcs), which means that if there is an arc $(w_i \rightarrow w_j)$, then w_i is an ancestor of all the words between w_i and w_j. Let $\Phi(W)$ denote the set of all the directed, projective trees that span W.

From an input sentence W, one would like to be able to compute the best parse; that is, a projective tree, $T \in \Phi(W)$, that obtains the highest "score". In particular, I follow Eisner (1996) and McDonald et al. (2005) and assume that the score of a complete spanning tree T for a given sentence, whether probabilistically motivated or not, can be decomposed as a sum of local scores for each link (a word pair). In which case, the parsing problem reduces to

$$T^* = \arg \max_{T \in \Phi(W)} \sum_{(w_i \rightarrow w_j) \in T} s(w_i \rightarrow w_j) \quad (1)$$

where the score $s(w_i \rightarrow w_j)$ can depend on any measurable property of w_i and w_j within the tree T. This formulation is sufficiently general to capture most dependency parsing models, including probabilistic dependency models (Wang et al., 2005; Eisner, 1996) as well as non-probabilistic models (McDonald et al., 2005; Wang et al., 2006).

For the purpose of learning, the score of each link can be expressed as a weighted linear combination of features

$$s(w_i \rightarrow w_j) = \boldsymbol{\theta}^\top \mathbf{f}(w_i \rightarrow w_j) \quad (2)$$

where $\boldsymbol{\theta}$ are the weight parameters to be estimated during training.

3 Lexicalised Dependency Parsing

To learn an accurate dependency parser from data, the first approach I investigated is based on a strictly lexical parsing model where all the parameters are based on words (Wang et al., 2005). The advantage of this approach is that it does not rely on part-of-speech tags nor grammatical categories. Furthermore, I based training on maximizing the *conditional* probability of a parse tree given a sentence, unlike most previous generative models (Magerman, 1995; Collins, 1997; Charniak, 1997), which focus on maximizing the joint probability of the parse tree and the sentence.

An efficient training algorithm can be achieved by maximizing the conditional probability of each parsing decision, hence minimizing a loss based on each local link decision independently. Importantly, inter-dependence between links can still be accommodated by exploiting *dynamic features* in training—features that take into account the *labels* of (some) of the surrounding components when predicting the label of a target component. To cope with the sparse data problem, I use distributional word similarity (Pereira et al., 1993; Grefenstette, 1994; Lin, 1998) to generalize the observed frequency counts in the training corpus. The experimental results on the Chinese Treebank 4.0 show that the accuracy of the conditional model is 13.6% higher than corresponding joint models, while similarity smoothing also allows the strictly lexicalised approach to outperform corresponding models based on part-of-speech tags.

4 Extensions to Large Margin Parsing

The approach presented above has a limitation: it uses a local scoring function instead of a global scoring function to compute the score for a candidate tree. The structured large margin approach, on the other hand, uses a global scoring function by minimizing a training loss—the "structured margin loss" (McDonald et al., 2005)—which is directly coordinated with the global tree. However, the training error minimized in the large margin approach is a coarse measure that only assesses the total error of an entire parse rather than focusing on the error of any particular component. Also, smoothing methods, which have been widely used in probabilistic approaches, are not currently being used in large margin training algorithms. In the second approach, I improve structured large margin training for parsing in two ways (Wang et al., 2006). First, I incorporate local constraints that enforce the correctness of each individual link, rather than just scoring the global parse tree. Second, to cope with sparse data and generalize to unseen words, I smooth the lexical parameters according to their underlying word similarities. To smooth parameters in the large margin framework, I introduce the technique of Laplacian

regularization in large margin parsing. Finally, to demonstrate the benefits of my approach, I reconsider the problem of parsing Chinese treebank data using only lexical features, as in Section 3. My results improve current large margin approaches and show that similarity smoothing combined with local constraint enforcement leads to state of the art performance, while only requiring word-based features that do not rely on part-of-speech tags nor grammatical categories in any way.

5 Training via Structured Boosting

Finally, I have recently demonstrated the somewhat surprising result that state of the art dependency parsing performance can be achieved through the use of conventional, local classification methods. In particular, I show how a simple form of structured boosting can be used to improve the training of standard local classification methods, in the context of structured predictions, without modifying the underlying training method (Wang et al., 2007). The advantage of this approach is that one can use off-the-shelf classification techniques, such as support vector machines or logistic regression, to achieve competitive parsing results with little additional effort.

The idea behind structured boosting is very simple. To produce an accurate parsing model, one combines the local predictions of multiple weak predictors to obtain a score for each link, which a parser can then use to compute the maximum score tree for a given sentence. Structured boosting proceeds in rounds. On each round a local "link predictor" is trained merely to predict the existence and orientation of a link between two words given input features encoding context—without worrying about coordinating the predictions in a coherent global parse. Once a weak predictor is learned, it is added to the ensemble of weak hypotheses, the training corpus is re-parsed using the new predictor, and the local training contexts are re-weighted based on errors made by the *parser*'s output. Thus, a wrapper approach is used to successively modify the training data so that the training algorithm is encouraged to facilitate improved global parsing accuracy.

Table 1: Comparison with State of the Art (Dependency Accuracy)

Model	Chinese	English
Yamada&Matsumoto 03	-	90.3
Nivre&Scholz 04	-	87.3
Wang et al. 05 (Sec. 3)	79.9*	-
McDonald et al. 05	-	90.9
McDonald&Pereira 06	82.5*	91.5
Corston-Oliver et al. 06	73.3†	90.8
Structured Boosting (Sec. 5)	86.6* 77.6†	89.3

* Obtained with Chinese Treebank 4.0 using the data split reported in Wang et al. (2005).
† Obtained with Chinese Treebank 5.0 using the data split reported in Corston-Olivr et al. (2006).

6 Current Results

Table 1 compares my results[1] with those obtained by other researchers, on both English and Chinese data.[2] The English results are obtained using the same standard training and test set splits from English Penn Treebank 3.0. The results on Chinese are obtained on two different data sets, Chinese Treebank 4.0 and Chinese Treebank 5.0 as noted.[3]

Table 1 shows that the results I am able to achieve on English are competitive with the state of the art, but are still behind the best results of (McDonald and Pereira, 2006). However, perhaps surprisingly, Table 1 also shows that the structured boosting approach actually surpasses state of the art accuracy on Chinese parsing for both treebank collections.

7 Future Work

Although the three pieces of my work above look very different superficially, they are actually closely related by the "scoring" formulation and, more

[1] I did not include the results of the technique described in Section 4, because we were only able to conveniently train on sentences with less than or equal to 15 words.

[2] McDonald et al. (2005) have tried MIRA on Chinese Treebank 4.0 with the same data split reported here, obtaining a dependency accuracy score of 82.5 (Ryan McDonald, personal communication).

[3] The results on Chinese Treebank 5.0 are generally worse than on Chinese Treebank 4.0, since the former is a superset of the latter, and moreover the additional sentences come entirely from a Taiwanese Chinese source that is more difficult to parse than the rest of the data.

specifically, by the equations introduced in Section 2. In other words, they all compute a linear classifier.[4] The only differences among them are: (1) What features are used? (2) How are the parameters $\boldsymbol{\theta}$ estimated?

A general perspective I bring to my investigation is the desire to delineate the effects of domain engineering (choosing good features for representing and learning parsing models) from the general machine learning principles (training criteria, regularization and smoothing techniques) that permit good results. In fact, combined features have been proved to be useful in dependency parsing with support vector machines (Yamada and Matsumoto, 2003), and I have already obtained some preliminary results on generating useful feature combinations via boosting. Therefore, I will consider combining all the projects I presented above. That is, I plan to incorporate all the useful features, the morphological features and the combined features as discussed above, into the training algorithms presented in Section 4 or Section 5, to train a dependency parser globally. Then I am going to augment the training with the existing smoothing and regularization techniques (as described in Section 4), or new developed ones. I expect the resulting parser to have better performance than those I have presented above.

There are a lot of other ideas which can be explored in my future work. First and most important, I plan to investigate new advanced machine learning methods (e.g., structured boosting or unsupervised / semi-supervised algorithms (Xu et al., 2006)) and apply them to the dependency parsing problem generally, since the goal of my research is to learn natural language parsers in an elegant and principled manner. Next, I am going to apply my approaches to parse other languages, such as Czech, German, Spanish and French, and analyze the performance of my parsers on these different languages. Furthermore, I plan to apply my parsers in other domains (e.g., biomedical data) (Blitzer et al., 2006) besides treebank data, to investigate the effectiveness and generality of my approaches.

[4] In general, for any probabilistic model, the product of probabilities can be converted to sums of scores in the log space, which makes the search identical to a score based discriminative model.

References

D. Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4).

J. Blitzer, R. McDonald, and F. Pereira. 2006. Doman adaptation with structural correspondence learning. In *Proc. of EMNLP*.

E. Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proc. of AAAI*, pages 598–603.

E. Charniak. 2000. A maximum entropy inspired parser. In *Proc. of North American ACL*, pages 132–139.

M. Collins. 1997. Three generative, lexicalized models for statistical parsing. In *Proc. of ACL*, pages 16–23.

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

S. Corston-Oliver, A. Aue, K. Duh, and E. Ringger. 2006. Multilingual dependency parsing using Bayes' point machines. In *Proc. of HLT/NAACL*.

J. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proc. of COLING*.

G. Grefenstette. 1994. Corpus-derived first, second and third-order word affinities. In *Proc. of Euralex*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING/ACL*, pages 768–774.

D. Magerman. 1995. Statistical decision-tree model for parsing. In *Proc. of ACL*, pages 276–283.

R. McDonald and F. Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proc. of EACL*.

R. McDonald, K. Crammer, and F. Pereira. 2005. Online large-margin training of dependency parsers. In *Proc. of ACL*.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of English words. In *Proc. of ACL*, pages 183–190.

A. Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. *Mach. Learn.*, 34(1-3):151–175.

B. Taskar, D. Klein, M. Collins, D. Koller, and C. Manning. 2004. Max-margin parsing. In *Proc. of EMNLP*.

Q. Wang, D. Schuurmans, and D. Lin. 2005. Strictly lexical dependency parsing. In *Proc. of IWPT*, pages 152–159.

Q. Wang, C. Cherry, D. Lizotte, and D. Schuurmans. 2006. Improved large margin dependency parsing via local constraints and Laplacian regularization. In *Proc. of CoNLL*.

Q. Wang, D. Lin, and D. Schuurmans. 2007. Simple training of dependency parsers via structured boosting. In *Proc. of IJCAI*, pages 1756–1762.

L. Xu, D. Wilkinson, F. Southey, and D. Schuurmans. 2006. Discriminative unsupervised learning of structured predictors. In *Proc. of ICML*.

H. Yamada and Y. Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proc. of IWPT*.