Fast and Robust Multilingual Dependency Parsing with a Generative Latent Variable Model

Ivan Titov
University of Geneva
24, rue Général Dufour
CH-1211 Genève 4, Switzerland
ivan.titov@cui.unige.ch

James Henderson
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, United Kingdom
james.henderson@ed.ac.uk

Abstract

We use a generative history-based model to predict the most likely derivation of a dependency parse. Our probabilistic model is based on Incremental Sigmoid Belief Networks, a recently proposed class of latent variable models for structure prediction. Their ability to automatically induce features results in multilingual parsing which is robust enough to achieve accuracy well above the average for each individual language in the multilingual track of the CoNLL-2007 shared task. This robustness led to the third best overall average labeled attachment score in the task, despite using no discriminative methods. We also demonstrate that the parser is quite fast, and can provide even faster parsing times without much loss of accuracy.

1 Introduction

The multilingual track of the CoNLL-2007 shared task (Nivre et al., 2007) considers dependency parsing of texts written in different languages. It requires use of a single dependency parsing model for the entire set of languages; model parameters are estimated individually for each language on the basis of provided training sets. We use a recently proposed dependency parser (Titov and Henderson, 2007b)[1] which has demonstrated state-of-the-art performance on a selection of languages from the CoNLL-X shared task (Buchholz and Marsi, 2006). This parser employs a latent variable model, Incremental Sigmoid Belief Networks (ISBNs), to define a generative history-based model of projective parsing. We used the pseudo-projective transformation introduced in (Nivre and Nilsson, 2005) to cast non-projective parsing tasks as projective. Following (Nivre et al., 2006), the encoding scheme called *HEAD* in (Nivre and Nilsson, 2005) was used to encode the original non-projective dependencies in the labels of the projectivized dependency tree. In the following sections we will briefly discuss our modifications to the ISBN parser, experimental setup, and achieved results.

2 The Probability Model

Our probability model uses the parsing order proposed in (Nivre et al., 2004), but instead of performing deterministic parsing as in (Nivre et al., 2004), this ordering is used to define a generative history-based model, by adding word prediction to the *Shift* parser action. We also decomposed some parser actions into sub-sequences of decisions. We split arc prediction decisions (*Left-Arc*$_r$ and *Right-Arc*$_r$) each into two elementary decisions: first the parser creates the corresponding arc, then it assigns a relation r to the arc. Similarly, we decompose the decision to shift a word into a decision to shift and a prediction of the word. We used part-of-speech tags and fine-grain word features, which are given in the data, to further decompose word predictions. First we predict the fine-grain part-of-speech tag for the word, then the set of word features (treating each set as an atomic value), and only then the particu-

[1] The ISBN parser will be soon made downloadable from the authors' web-page.

Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, pp. 947–951,
Prague, June 2007. ©2007 Association for Computational Linguistics

lar word form. This approach allows us to both decrease the effect of sparsity and to avoid normalization across all the words in the vocabulary, significantly reducing the computational expense of word prediction. When conditioning on words, we treated each word feature individually, as this proved to be useful in (Titov and Henderson, 2007b).

The probability of each parser decision, conditioned on the complete parse history, is modeled using a form a graphical model called Incremental Sigmoid Belief Networks. ISBNs, originally proposed for constituent parsing in (Titov and Henderson, 2007a), use vectors of binary latent variables to encode information about the parse history. These history variables are similar to the hidden state of a Hidden Markov Model. But unlike the graphical model for an HMM, which would specify conditional dependency edges only between adjacent states in the parse history, the ISBN graphical model can specify conditional dependency edges between latent variables which are arbitrarily far apart in the parse history. The source state of such an edge is determined by the partial parse structure built at the time of the destination state, thereby allowing the conditional dependency edges to be appropriate for the structural nature of the parsing problem. In particular, they allow conditional dependencies to be local in the parse structure, not just local in the history sequence. In this they are similar to the class of neural networks proposed in (Henderson, 2003) for constituent parsing. In fact, in (Titov and Henderson, 2007a) it was shown that this neural network can be viewed as a coarse approximation to the corresponding ISBN model.

Traditional statistical parsing models also condition on features which are local in the parse structure, but these features need to be explicitly defined before learning, and require careful feature selection. This is especially difficult for languages unknown to the parser developer, since the number of possible features grows exponentially with the structural distance considered.

The ISBN model uses an alternative approach, where latent variables are used to induce features during learning. The most important problem in designing an ISBN is to define an appropriate structural locality for each parser decision. This is done by choosing a fixed set of relationships between parser states, where the information which is needed to make the decision at the earlier state is also useful in making the decision at the later state. The latent variables for these related states are then connected with conditional dependency edges in the ISBN graphical model. Longer conditional dependencies are then possible through chains of these immediate conditional dependencies, but there is an inductive bias toward shorter chains. This bias makes it important that the set of chosen relationships defines an appropriate notion of locality. However, as long as there exists some chain of relationships between any two states, then any statistical dependency which is clearly manifested in the data can be learned, even if it was not foreseen by the designer. This provides a potentially powerful form of feature induction, which is nonetheless biased toward a notion of locality appropriate for the nature of the problem.

In our experiments we use the same definition of structural locality as was proposed for the ISBN dependency parser in (Titov and Henderson, 2007b). The current state is connected to previous states using a set of 7 distinct relationships defined in terms of each state's parser configuration, which includes of a stack and a queue. Specifically, the current state is related to the last previous state whose parser configuration has: the same queue, the same stack, a stack top which is the rightmost right child of the current stack top, a stack top which is the leftmost left child of the current stack top, a front of the queue which is the leftmost child of the front of the current queue, a stack top which is the head word of the current stack top, a front of the queue which is the current stack top. Different model parameters are trained for each of these 7 types of relationship, but the same parameters are used everywhere in the graphical model where the relationship holds.

Each latent variable in the ISBN parser is also conditionally dependent on a set of explicit features of the parsing history. As long as these explicit features include all the new information from the last parser decision, the performance of the model is not very sensitive to this design choice. We used the base feature model defined in (Nivre et al., 2006) for all the languages but Arabic, Chinese, Czech, and Turkish. For Arabic, Chinese, and Czech, we used the same feature models used in the CoNLL-X shared task by (Nivre et al., 2006), and for Turkish we used again the base feature model but extended it with a single feature: the part-of-speech tag of the token preceding the current top of the stack.

3 Parsing

Exact inference in ISBN models is not tractable, but effective approximations were proposed in (Titov and Henderson, 2007a). Unlike (Titov and Henderson, 2007b), in the shared task we used only the simplest feed-forward approximation, which replicates the computation of a neural network of the type proposed in (Henderson, 2003). We would expect better performance with the more accurate approximation based on variational inference proposed and evaluated in (Titov and Henderson, 2007a). We did not try this because, on larger treebanks it would have taken too long to tune the model with this better approximation, and using different approximation methods for different languages would not be compatible with the shared task rules.

To search for the most probable parse, we use the heuristic search algorithm described in (Titov and Henderson, 2007b), which is a form of beam search. In section 4 we show that this search leads to quite efficient parsing.

To overcome a minor shortcoming of the parsing algorithm of (Nivre et al., 2004) we introduce a simple language independent post-processing step. Nivre's parsing algorithm allows unattached nodes to stay on the stack at the end of parsing, which is reasonable for treebanks with unlabeled attachment to root. However, this sometimes happens with languages where only labeled attachment to root is allowed. In these cases (only 35 tokens in Greek, 17 in Czech, 1 in Arabic, on the final testing set) we attached them using a simple rule: if there are no tokens in the sentence attached to root, then the considered token is attached to root with the most frequent root-attachment relation used for its part-of-speech tag. If there are other root-attached tokens in the sentence, it is attached to the next root-attached token with the most frequent relation. Preference is given to the most frequent attachment direction for its part-of-speech tag. This rule guarantees that no loops are introduced by the post-processing.

4 Experiments

We evaluated the ISBN parser on all the languages considered in the shared task (Hajič et al., 2004; Aduriz et al., 2003; Martí et al., 2007; Chen et al., 2003; Böhmová et al., 2003; Marcus et al., 1993; Johansson and Nugues, 2007; Prokopidis et al., 2005; Csendes et al., 2005; Montemagni et al., 2003; Oflazer et al., 2003). ISBN models were trained using a small development set taken out from the training set, which was used for tuning learning and decoding parameters, for early stopping and very coarse feature engineering.[2] The sizes of the development sets were different: starting from less than 2,000 tokens for smaller treebanks to 5,000 tokens for the largest one. The relatively small sizes of the development sets limited our ability to perform careful feature selection, but this should not have significantly affected the model performance, as discussed in section 2.[3] We used frequency cutoffs: we ignored any property (word form, lemma, feature) which occurs in the training set less than a given threshold. We used a threshold of 20 for Greek and Chinese and a threshold of 5 for the rest. Because cardinalities of each of these sets (sets of word forms, lemmas and features) effect the model efficiency, we selected the larger threshold when validation results with the smaller threshold were comparable. For the ISBN latent variables, we used vectors of length 80, based on our previous experience.

Results on the final testing set are presented in table 1. The model achieves relatively high scores on each individual language, significantly better than each average result in the shared task. This leads to the third best overall average results in the shared task, both in average labeled attachment score and in average unlabeled attachment score. The absolute error increase in labeled attachment score over the best system is only 0.4%. We attribute ISBN's success mainly to its ability to automatically induce features, as this significantly reduces the risk of omitting any important highly predictive features. This makes an ISBN parser a particularly good baseline when considering a new treebank or language, be-

[2] We plan to make all the learning and decoding parameters available on our web-page.

[3] Use of cross-validation with our model is relatively time-consuming and, thus, not quite feasible for the shared task.

	Ara	Bas	Cat	Chi	Cze	Eng	Gre	Hun	Ita	Tur	**Ave**
LAS	74.1	75.5	87.4	82.1	77.9	88.4	73.5	77.9	82.3	79.8	**79.90**
UAS	83.2	81.9	93.4	87.9	84.2	89.7	81.2	82.2	86.3	86.2	**85.62**

Table 1: Labeled attachment score (LAS) and unlabeled attachment score (UAS) on the final testing sets

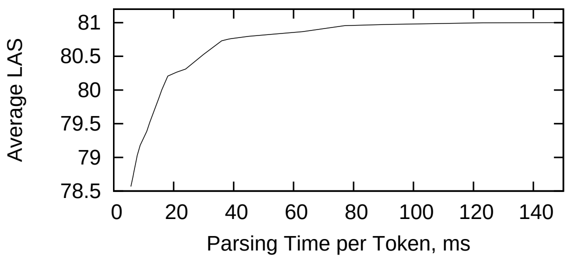

Figure 1: Average labeled attachment score on Basque, Chinese, English, and Turkish development sets as a function of parsing time per token

cause it does not require much effort in feature engineering. As was demonstrated in (Titov and Henderson, 2007b), even a minimal set of local explicit features achieves results which are non-significantly different from a carefully chosen set of explicit features, given the language independent definition of locality described in section 2.

It is also important to note that the model is quite efficient. Figure 1 shows the tradeoff between accuracy and parsing time as the width of the search beam is varied, on the development set. This curve plots the average labeled attachment score over Basque, Chinese, English, and Turkish as a function of parsing time per token.[4] Accuracy of only 1% below the maximum can be achieved with average processing time of 17 ms per token, or 60 tokens per second.[5]

We also refer the reader to (Titov and Henderson, 2007b) for more detailed analysis of the ISBN dependency parser results, where, among other things, it was shown that the ISBN model is especially accurate at modeling long dependencies.

5 Conclusion

We evaluated the ISBN dependency parser in the multilingual shared task setup and achieved competitive accuracy on every language, and the third best average score overall. The proposed model requires minimal design effort because it relies mostly on automatic feature induction, which is highly desirable when using new treebanks or languages. The parsing time needed to achieve high accuracy is also quite small, making this model a good candidate for use in practical applications.

The fact that our model defines a probability model over parse trees, unlike the previous state-of-the-art methods (Nivre et al., 2006; McDonald et al., 2006), makes it easier to use this model in applications which require probability estimates, such as in language processing pipelines or for language modeling. Also, as with any generative model, it should be easy to improve the parser's accuracy with discriminative reranking, such as discriminative retraining techniques (Henderson, 2004) or data-defined kernels (Henderson and Titov, 2005), with or even without the introduction of any additional linguistic features.

Acknowledgments

This work was funded by Swiss NSF grant 200020-109685, UK EPSRC grant EP/E019501/1, and EU FP6 grant 507802 for project TALK.

References

A. Abeillé, editor. 2003. *Treebanks: Building and Using Parsed Corpora*. Kluwer.

I. Aduriz, M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Diaz de Ilarraza, A. Garmendia, and M. Oronoz. 2003. Construction of a Basque dependency treebank. In *Proc. of the 2nd Workshop on Treebanks and Linguistic Theories (TLT)*, pages 201–204.

A. Böhmová, J. Hajič, E. Hajičová, and B. Hladká. 2003. The PDT: a 3-level annotation scenario. In Abeillé (Abeillé, 2003), chapter 7, pages 103–127.

[4] A piecewise-linear approximation for each individual language was used to compute the average. Experiments were run on a standard 2.4 GHz desktop PC.

[5] For Basque, Chinese, and Turkish this time is below 7 ms, but for English it is 38 ms. English, along with Catalan, required the largest beam across all 10 languages. Note that accuracy in the lowest part of the curve can probably be improved by varying latent vector size and frequency cut-offs. Also, efficiency was not the main goal during the implementation of the parser, and it is likely that a much faster implementation is possible.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proc. of the Tenth Conference on Computational Natural Language Learning*, New York, USA.

K. Chen, C. Luo, M. Chang, F. Chen, C. Chen, C. Huang, and Z. Gao. 2003. Sinica treebank: Design criteria, representational issues and implementation. In Abeillé (Abeillé, 2003), chapter 13, pages 231–248.

D. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. 2005. *The Szeged Treebank.* Springer.

J. Hajič, O. Smrž, P. Zemánek, J. Šnaidauf, and E. Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, pages 110–117.

James Henderson and Ivan Titov. 2005. Data-defined kernels for parse reranking derived from probabilistic models. In *Proc. 43rd Meeting of Association for Computational Linguistics*, Ann Arbor, MI.

James Henderson. 2003. Inducing history representations for broad coverage statistical parsing. In *Proc. joint meeting of North American Chapter of the Association for Computational Linguistics and the Human Language Technology Conf.*, pages 103–110, Edmonton, Canada.

James Henderson. 2004. Discriminative training of a neural network statistical parser. In *Proc. 42nd Meeting of Association for Computational Linguistics*, Barcelona, Spain.

R. Johansson and P. Nugues. 2007. Extended constituent-to-dependency conversion for English. In *Proc. of the 16th Nordic Conference on Computational Linguistics (NODALIDA)*.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M. A. Martí, M. Taulé, L. Màrquez, and M. Bertran. 2007. CESS-ECE: A multilingual and multilevel annotated corpus. Available for download from: http://www.lsi.upc.edu/∼mbertran/cess-ece/.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proc. of the Tenth Conference on Computational Natural Language Learning*, New York, USA.

S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Nana, F. Pianesi, and R. Delmonte. 2003. Building the Italian Syntactic-Semantic Treebank. In Abeillé (Abeillé, 2003), chapter 11, pages 189–210.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proc. 43rd Meeting of Association for Computational Linguistics*, pages 99–106, Ann Arbor, MI.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proc. of the Eighth Conference on Computational Natural Language Learning*, pages 49–56, Boston, USA.

Joakim Nivre, Johan Hall, Jens Nilsson, Gulsen Eryigit, and Svetoslav Marinov. 2006. Pseudo-projective dependency parsing with support vector machines. In *Proc. of the Tenth Conference on Computational Natural Language Learning*, pages 221–225, New York, USA.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proc. of the CoNLL 2007 Shared Task. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

K. Oflazer, B. Say, D. Zeynep Hakkani-T ür, and G. Tür. 2003. Building a Turkish treebank. In Abeillé (Abeillé, 2003), chapter 15, pages 261–277.

P. Prokopidis, E. Desypri, M. Koutsombogera, H. Papageorgiou, and S. Piperidis. 2005. Theoretical and practical issues in the construction of a Greek dependency treebank. In *Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 149–160.

Ivan Titov and James Henderson. 2007a. Constituent parsing with incremental sigmoid belief networks. In *Proc. 45th Meeting of Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Ivan Titov and James Henderson. 2007b. A latent variable model for generative dependency parsing. In *Proc. 10th Int. Conference on Parsing Technologies (IWPT)*, Prague, Czech Republic.