A Latent Variable Model for Generative Dependency Parsing

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 propose a generative dependency parsing model which uses binary latent variables to induce conditioning features. To define this model we use a recently proposed class of Bayesian Networks for structured prediction, Incremental Sigmoid Belief Networks. We demonstrate that the proposed model achieves state-of-the-art results on three different languages. We also demonstrate that the features induced by the ISBN's latent variables are crucial to this success, and show that the proposed model is particularly good on long dependencies.

1 Introduction

Dependency parsing has been a topic of active research in natural language processing during the last several years. The CoNLL-X shared task (Buchholz and Marsi, 2006) made a wide selection of standardized treebanks for different languages available for the research community and allowed for easy comparison between various statistical methods on a standardized benchmark. One of the surprising things discovered by this evaluation is that the best results are achieved by methods which are quite different from state-of-the-art models for constituent parsing, e.g. the deterministic parsing method of (Nivre et al., 2006) and the minimum spanning tree parser of (McDonald et al., 2006). All the most accurate dependency parsing models are fully discriminative, unlike constituent parsing where all the state of the art methods have a generative component (Charniak and Johnson, 2005; Henderson, 2004; Collins, 2000). Another surprising thing is the lack of latent variable models among the methods used in the shared task. Latent variable models would allow complex features to be induced automatically, which would be highly desirable in multilingual parsing, where manual feature selection might be very difficult and time consuming, especially for languages unknown to the parser developer.

In this paper we propose a generative latent variable model for dependency parsing. It is based on Incremental Sigmoid Belief Networks (ISBNs), a class of directed graphical model for structure prediction problems recently proposed in (Titov and Henderson, 2007), where they were demonstrated to achieve competitive results on the constituent parsing task. As discussed in (Titov and Henderson, 2007), computing the conditional probabilities which we need for parsing is in general intractable with ISBNs, but they can be approximated efficiently in several ways. In particular, the neural network constituent parsers in (Henderson, 2003) and (Henderson, 2004) can be regarded as coarse approximations to their corresponding ISBN model.

ISBNs use history-based probability models. The most common approach to handling the unbounded nature of the parse histories in these models is to choose a pre-defined set of features which can be unambiguously derived from the history (e.g. (Charniak, 2000; Collins, 1999; Nivre et al., 2004)). Decision probabilities are then assumed to be independent of all information not represented by this finite set of features. ISBNs instead use a vector of binary

Proceedings of the 10th Conference on Parsing Technologies, pages 144–155,
Prague, Czech Republic, June 2007.

latent variables to encode the information about the parser history. This history vector is similar to the hidden state of a Hidden Markov Model. But unlike the graphical model for an HMM, which specifies conditional dependency edges only between adjacent states in the sequence, the ISBN graphical model can specify conditional dependency edges between states which are arbitrarily far apart in the parse history. The source state of such an edge is determined by the partial output structure built at the time of the destination state, thereby allowing the conditional dependency edges to be appropriate for the structural nature of the problem being modeled. This structure sensitivity is possible because ISBNs are a constrained form of switching model (Murphy, 2002), where each output decision switches the model structure used for the remaining decisions.

We build an ISBN model of dependency parsing using the parsing order proposed in (Nivre et al., 2004). However, instead of performing deterministic parsing as in (Nivre et al., 2004), we use this ordering to define a generative history-based model, by integrating word prediction operations into the set of parser actions. Then we propose a simple, language independent set of relations which determine how latent variable vectors are interconnected by conditional dependency edges in the ISBN model. ISBNs also condition the latent variable vectors on a set of explicit features, which we vary in the experiments.

In experiments we evaluate both the performance of the ISBN dependency parser compared to previous work, and the ability of the ISBN model to induce complex history features. Our model achieves state-of-the-art performance on the languages we test, significantly outperforming the model of (Nivre et al., 2006) on two languages out of three and demonstrating about the same results on the third. In order to test the model's feature induction abilities, we train models with two different sets of explicit conditioning features: the feature set individually tuned by (Nivre et al., 2006) for each considered language, and a minimal set of local features. These models achieve comparable accuracy, unlike with the discriminative SVM-based approach of (Nivre et al., 2006), where careful feature selection appears to be crucial. We also conduct a controlled experiment where we used the tuned features of (Nivre et al., 2006) but disable the feature induction abilities of our model by elimination of the edges connecting latent state vectors. This restricted model achieves far worse results, showing that it is exactly the capacity of ISBNs to induce history features which is the key to its success. It also motivates further research into how feature induction techniques can be exploited in discriminative parsing methods.

We analyze how the relation accuracy changes with the length of the head-dependent relation, demonstrating that our model very significantly outperforms the state-of-the-art baseline of (Nivre et al., 2006) on long dependencies. Additional experiments suggest that both feature induction abilities and use of the beam search contribute to this improvement.

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, e.g. in language processing pipelines. Also, as with any generative model, it may be easy to improve the parser's accuracy by using discriminative retraining techniques (Henderson, 2004) or data-defined kernels (Henderson and Titov, 2005), with or even without introduction of any additional linguistic features. In addition, there are some applications, such as language modeling, which require generative models. Another advantage of generative models is that they do not suffer from the label bias problems (Bottou, 1991), which is a potential problem for conditional or deterministic history-based models, such as (Nivre et al., 2004).

In the remainder of this paper, we will first review general ISBNs and how they can be approximated. Then we will define the generative parsing model, based on the algorithm of (Nivre et al., 2004), and propose an ISBN for this model. The empirical part of the paper then evaluates both the overall accuracy of this method and the importance of the model's capacity to induce features. Additional related work will be discussed in the last section before concluding.

2 The Latent Variable Architecture

In this section we will begin by briefly introducing the class of graphical models we will be using, Incremental Sigmoid Belief Networks (Titov and Henderson, 2007). ISBNs are designed specifically for modeling structured data. They are latent variable models which are not tractable to compute exactly, but two approximations exist which have been shown to be effective for constituent parsing (Titov and Henderson, 2007). Finally, we present how these approximations can be trained.

2.1 Incremental Sigmoid Belief Networks

An ISBN is a form of Sigmoid Belief Network (SBN) (Neal, 1992). SBNs are Bayesian Networks with binary variables and conditional probability distributions in the form:

$$P(S_i = 1|Par(S_i)) = \sigma(\sum_{S_j \in Par(S_i)} J_{ij} S_j),$$

where S_i are the variables, $Par(S_i)$ are the variables which S_i depends on (its parents), σ denotes the logistic sigmoid function, and J_{ij} is the weight for the edge from variable S_j to variable S_i in the graphical model. SBNs are similar to feed-forward neural networks, but unlike neural networks, SBNs have a precise probabilistic semantics for their hidden variables. ISBNs are based on a generalized version of SBNs where variables with any range of discrete values are allowed. The normalized exponential function ('soft-max') is used to define the conditional probability distributions at these nodes.

To extend SBNs for processing arbitrarily long sequences, such as a parser's sequence of decisions $D^1, ..., D^m$, SBNs are extended to a form of Dynamic Bayesian Network (DBN). In DBNs, a new set of variables is instantiated for each position in the sequence, but the edges and weights are the same for each position in the sequence. The edges which connect variables instantiated for different positions must be directed forward in the sequence, thereby allowing a temporal interpretation of the sequence.

Incremental Sigmoid Belief Networks (Titov and Henderson, 2007) differ from simple dynamic SBNs in that they allow the model structure to depend on the output variable values. Specifically, a decision is allowed to effect the placement of any edge whose destination is after the decision. This results in a form of switching model (Murphy, 2002), where each decision switches the model structure used for the remaining decisions. The incoming edges for a given position are a discrete function of the sequence of decisions which precede that position. This makes the ISBN an "incremental" model, not just a dynamic model. The structure of the model is determined incrementally as the decision sequence proceeds.

ISBNs are designed to allow the model structure to depend on the output values without overly complicating the inference of the desired conditional probabilities $P(D^t|D^1, \ldots, D^{t-1})$, the probability of the next decision given the history of previous decisions. In particular, it is never necessary to sum over all possible model structures, which in general would make inference intractable.

2.2 Modeling Structures with ISBNs

ISBNs are designed for modeling structured data where the output structure is not given as part of the input. In dependency parsing, this means they can model the probability of an output dependency structure when the input only specifies the sequence of words (i.e. parsing). The difficulty with such problems is that the statistical dependencies in the dependency structure are local in the structure, and not necessarily local in the word sequence. ISBNs allow us to capture these statistical dependencies in the model structure by having model edges depend on the output variables which specify the dependency structure. For example, if an output specifies that there is a dependency arc from word w_i to word w_j, then any future decision involving w_j can directly depend on its head w_i. This allows the head w_i to be treated as local to the dependent w_j even if they are far apart in the sentence.

This structurally-defined notion of locality is particularly important for the model's latent variables. When the structurally-defined model edges connect latent variables, information can be propagated between latent variables, thereby providing an even larger structural domain of locality than that provided by single edges. This provides a potentially powerful form of feature induction, which is nonetheless biased toward a notion of locality which is appropriate for the structure of the problem.

2.3 Approximating ISBNs

(Titov and Henderson, 2007) proposes two approximations for inference in ISBNs, both based on variational methods. The main idea of variational methods (Jordan et al., 1999) is, roughly, to construct a tractable approximate model with a number of free parameters. The values of the free parameters are set so that the resulting approximate model is as close as possible to the original graphical model for a given inference problem.

The simplest example of a variation method is the mean field method, which uses a fully factorized distribution $Q(H|V) = \prod_i Q_i(h_i|V)$ as the approximate model, where V are the visible (i.e. known) variables, $H = h_1, \ldots, h_l$ are the hidden (i.e. latent) variables, and each Q_i is the distribution of an individual latent variable h_i. The free parameters of this approximate model are the means μ_i of the distributions Q_i.

(Titov and Henderson, 2007) proposes two approximate models based on the variational approach. First, they show that the neural network of (Henderson, 2003) can be viewed as a coarse mean field approximation of ISBNs, which they call the feed-forward approximation. This approximation imposes the constraint that the free parameters μ_i of the approximate model are only allowed to depend on the distributions of their parent variables. This constraint increases the potential for a large approximation error, but it significantly simplifies the computations by allowing all the free parameters to be set in a single pass over the model.

The second approximation proposed in (Titov and Henderson, 2007) takes into consideration the fact that, after each decision is made, all the preceding latent variables should have their means μ_i updated. This approximation extends the feed-forward approximation to account for the most important components of this update. They call this approximation the mean field approximation, because a mean field approximation is applied to handle the statistical dependencies introduced by the new decisions. This approximation was shown to be a more accurate approximation of ISBNs than the feed-forward approximation, but remain tractable. It was also shown to achieve significantly better accuracy on constituent parsing.

2.4 Learning

Training these approximations of ISBNs is done to maximize the fit of the *approximate* models to the data. We use gradient descent, and a regularized maximum likelihood objective function. Gaussian regularization is applied, which is equivalent to the weight decay standardly used in neural networks. Regularization was reduced through the course of learning.

Gradient descent requires computing the derivatives of the objective function with respect to the model parameters. In the feed-forward approximation, this can be done with the standard Backpropagation learning used with neural networks. For the mean field approximation, propagating the error all the way back through the structure of the graphical model requires a more complicated calculation, but it can still be done efficiently (see (Titov and Henderson, 2007) for details).

3 The Dependency Parsing Algorithm

The sequences of decisions $D^1, ..., D^m$ which we will be modeling with ISBNs are the sequences of decisions made by a dependency parser. For this we use the parsing strategy for projective dependency parsing introduced in (Nivre et al., 2004), which is similar to a standard shift-reduce algorithm for context-free grammars (Aho et al., 1986). It can be viewed as a mixture of bottom-up and top-down parsing strategies, where left dependencies are constructed in a bottom-up fashion and right dependencies are constructed top-down. For details we refer the reader to (Nivre et al., 2004). In this section we briefly describe the algorithm and explain how we use it to define our history-based probability model.

In this paper, as in the CoNLL-X shared task, we consider labeled dependency parsing. The state of the parser is defined by the current stack S, the queue I of remaining input words and the partial labeled dependency structure constructed by previous parser decisions. The parser starts with an empty stack S and terminates when it reaches a configuration with an empty queue I. The algorithm uses 4 types of decisions:

1. The decision **Left-Arc**$_r$ adds a dependency arc from the next input word w_j to the word w_i on top of the stack and selects the label r for the

relation between w_i and w_j. Word w_i is then popped from the stack.

2. The decision **Right-Arc**$_r$ adds an arc from the word w_i on top of the stack to the next input word w_j and selects the label r for the relation between w_i and w_j.

3. The decision **Reduce** pops the word w_i from the stack.

4. The decision **Shift**$_{w_j}$ shifts the word w_j from the queue to the stack.

Unlike the original definition in (Nivre et al., 2004) the **Right-Arc**$_r$ decision does not shift w_j to the stack. However, the only thing the parser can do after a **Right-Arc**$_r$ decision is to choose the **Shift**$_{w_j}$ decision. This subtle modification does not change the actual parsing order, but it does simplify the definition of our graphical model, as explained in section 4.

We use a history-based probability model, which decomposes the probability of the parse according to the parser decisions:

$$P(T) = P(D^1, ..., D^m) = \prod_t P(D^t|D^1, \ldots, D^{t-1}),$$

where T is the parse tree and $D^1, \ldots, D^m$ is its equivalent sequence of parser decisions. Since we need a generative model, the action **Shift**$_{w_j}$ also predicts the next word in the queue I, w_{j+1}, thus the $P(Shift_{w_i}|D^1, \ldots, D^{t-1})$ is a probability both of the shift operation and the word w_{j+1} conditioned on current parsing history.[1]

Instead of treating each D^t as an atomic decision, it is convenient to split it into a sequence of elementary decisions $D^t = d_1^t, \ldots, d_n^t$:

$$P(D^t|D^1, \ldots, D^{t-1}) = \prod_k P(d_k^t|h(t,k)),$$

[1]In preliminary experiments, we also considered look-ahead, where the word is predicted earlier than it appears at the head of the queue I, and "anti-look-ahead", where the word is predicted only when it is shifted to the stack S. Early prediction allows conditioning decision probabilities on the words in the look-ahead and, thus, speeds up the search for an optimal decision sequence. However, the loss of accuracy with look-ahead was quite significant. The described method, where a new word is predicted when it appears at the head of the queue, led to the most accurate model and quite efficient search. The anti-look-ahead model was both less accurate and slower.

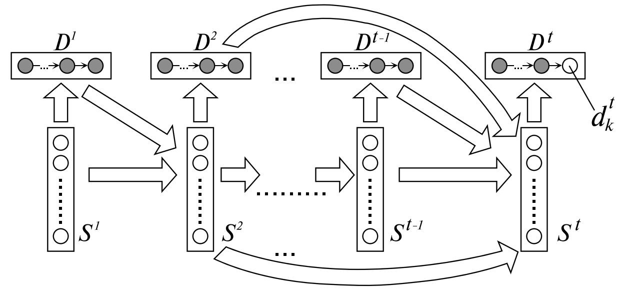

Figure 1: An ISBN for estimating $P(d_k^t|h(t,k))$.

where $h(t,k)$ denotes the parsing history $D^1, \ldots, D^{t-1}, d_1^t, \ldots, d_{k-1}^t$. We split **Left-Arc**$_r$ and **Right-Arc**$_r$ each into two elementary decisions: first, the parser decides to create the corresponding arc, then, it decides to assign a relation r to the arc. Similarly, we decompose the decision **Shift**$_{w_j}$ into an elementary decision to shift a word and a prediction of the word w_{j+1}. In our experiments we use datasets from the CoNLL-X shared task, which provide additional properties for each word token, such as its part-of-speech tag and some fine-grain features. This information implicitly induces word clustering, which we use in our model: first we predict a part-of-speech tag for the word, then a set of word features, treating feature combination as an atomic value, and only then a particular 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.

4 An ISBN for Dependency Parsing

In this section we define the ISBN model we use for dependency parsing. An example of this ISBN for estimating $P(d_k^t|h(t,k))$ is illustrated in figure 1. It is organized into vectors of variables: latent state variable vectors $S^{t'} = s_1^{t'}, \ldots, s_n^{t'}$, representing an intermediate state at position t', and decision variable vectors $D^{t'}$, representing a decision at position t', where $t' \leq t$. Variables whose value are given at the current decision (t,k) are shaded in figure 1, latent and current decision variables are left unshaded.

As illustrated by the edges in figure 1, the probability of each state variable $s_i^{t'}$ (the individual circles in $S^{t'}$) depends on all the variables in a finite set of relevant previous state and decision vectors, but there are no direct dependencies between the different variables in a single state vector. For each relevant decision vector, the precise set of decision variables which are connected in this way can be adapted to a particular language. As long as these connected decisions include all the new information about the parse, the performance of the model is not very sensitive to this choice. This is because ISBNs have the ability to induce their own complex features of the parse history, as demonstrated in the experiments in section 6.

The most important design decision in building an ISBN model is choosing the finite set of relevant previous state vectors for the current decision. By connecting to a previous state, we place that state in the local context of the current decision. This specification of the domain of locality determines the inductive bias of learning with ISBNs. When deciding what information to store in its latent variables, an ISBN is more likely to choose information which is immediately local to the current decision. This stored information then becomes local to any following connected decision, where it again has some chance of being chosen as relevant to that decision. In this way, the information available to a given decision can come from arbitrarily far away in the chain of interconnected states, but it is much more likely to come from a state which is relatively local. Thus, we need to choose the set of local (i.e. connected) states in accordance with our prior knowledge about which previous decisions are likely to be particularly relevant to the current decision.

To choose which previous decisions are particularly relevant to the current decision, we make use of the partial dependency structure which has been decided so far in the parse. Specifically, the current latent state vector is connected to a set of 7 previous latent state vectors (if they exist) according to the following relationships:

1. **Input Context**: the last previous state with the same queue I.
2. **Stack Context**: the last previous state with the same stack S.
3. **Right Child of Top of** S: the last previous state where the rightmost right child of the current stack top was on top of the stack.
4. **Left Child of Top of** S: the last previous state where the leftmost left child of the current stack top was on top of the stack.
5. **Left Child of Front of** I[2] : the last previous state where the leftmost child of the front element of I was on top of the stack.
6. **Head of Top**: the last previous state where the head word of the current stack top was on top of the stack.
7. **Top of** S **at Front of** I: the last previous state where the current stack top was at the front of the queue.

Each of these 7 relations has its own distinct weight matrix for the resulting edges in the ISBN, but the same weight matrix is used at each position where the relation is relevant.

All these relations but the last one are motivated by linguistic considerations. The current decision is primarily about what to do with the current word on the top of the stack and the current word on the front of the queue. The *Input Context* and *Stack Context* relationships connect to the most recent states used for making decisions about each of these words. The *Right Child of Top of S* relationship connects to a state used for making decisions about the most recently attached dependent of the stack top. Similarly, the *Left Child of Front of I* relationship connects to a state for the most recently attached dependent of the queue front. The *Left Child of Top of S* is the first dependent of the stack top, which is a particularly informative dependent for many languages. Likewise, the *Head of Top* can tell us a lot about the stack top, if it has been chosen already.

A second motivation for including a state in the local context of a decision is that it might contain information which has no other route for reaching the current decision. In particular, it is generally a good idea to ensure that the immediately preceding state is always included somewhere in the set of connected states. This requirement ensures that information, at least theoretically, can pass between any two states in the decision sequence, thereby avoiding any hard

[2]We refer to the *head* of the queue as the *front*, to avoid unnecessary ambiguity of the word *head* in the context of dependency parsing.

independence assumptions. The last relation, *Top of S at Front of I*, is included mainly to fulfill this requirement. Otherwise, after a **Shift**$_{w_j}$ operation, the preceding state would not be selected by any of the relationships.

As indicated in figure 1, the probability of each elementary decision $d_k^{t'}$ depends both on the current state vector $S^{t'}$ and on the previously chosen elementary action $d_{k-1}^{t'}$ from $D^{t'}$. This probability distribution has the form of a normalized exponential:

$$P(d_k^{t'} = d | S^{t'}, d_{k-1}^{t'}) = \frac{\Phi_{h(t',k)}\,(d)\, e^{\sum_j W_{dj} s_j^{t'}}}{\sum_{d'} \Phi_{h(t',k)}\,(d')\, e^{\sum_j W_{d'j} s_j^{t'}}},$$

where $\Phi_{h(t',k)}$ is the indicator function of the set of elementary decisions that may possibly follow the last decision in the history $h(t',k)$, and the W_{dj} are the weights. Now it is easy to see why the original decision **Right-Arc**$_r$ (Nivre et al., 2004) had to be decomposed into two distinct decisions: the decision to construct a labeled arc and the decision to shift the word. Use of this composite **Right-Arc**$_r$ would have required the introduction of individual parameters for each pair (w,r), where w is an arbitrary word in the lexicon and r - an arbitrary dependency relation.

5 Searching for the Best Tree

ISBNs define a probability model which does not make any a-priori assumptions of independence between any decision variables. As we discussed in section 4 use of relations based on partial output structure makes it possible to take into account statistical interdependencies between decisions closely related in the output structure, but separated by multiple decisions in the input structure. This property leads to exponential complexity of complete search. However, the success of the deterministic parsing strategy which uses the same parsing order (Nivre et al., 2006), suggests that it should be relatively easy to find an accurate approximation to the best parse with heuristic search methods. Unlike (Nivre et al., 2006), we can not use a lookahead in our generative model, as was discussed in section 3, so a greedy method is unlikely to lead to a good approximation. Instead we use a pruning strategy similar to that described in (Henderson, 2003), where it was applied to a considerably harder search problem: constituent parsing with a left-corner parsing order.

We apply fixed beam pruning after each decision **Shift**$_{w_j}$, because knowledge of the next word in the queue I helps distinguish unlikely decision sequences. We could have used best-first search between **Shift**$_{w_j}$ operations, but this still leads to relatively expensive computations, especially when the set of dependency relations is large. However, most of the word pairs can possibly participate only in a very limited number of distinct relations. Thus, we pursue only a fixed number of relations r after each **Left-Arc**$_r$ and **Right-Arc**$_r$ operation.

Experiments with a variety of post-shift beam widths confirmed that very small validation performance gains are achieved with widths larger than 30, and sometimes even a beam of 5 was sufficient. We found also that allowing 5 different relations after each dependency prediction operation was enough that it had virtually no effect on the validation accuracy.

6 Empirical Evaluation

In this section we evaluate the ISBN model for dependency parsing on three treebanks from the CoNLL-X shared task. We compare our generative models with the best parsers from the CoNLL-X task, including the SVM-based parser of (Nivre et al., 2006) (the MALT parser), which uses the same parsing algorithm. To test the feature induction abilities of our model we compare results with two feature sets, the feature set tuned individually for each language by (Nivre et al., 2006), and another feature set which includes only obvious local features. This simple feature set comprises only features of the word on top of the stack S and the front word of the queue I. We compare the gain from using tuned features with the similar gain obtained by the MALT parser. To obtain these results we train the MALT parser with the same two feature sets.[3]

In order to distinguish the contribution of ISBN's feature induction abilities from the contribution of

[3]The tuned feature sets were obtained from http://w3.msi.vxu.se/˜nivre/research/MaltParser.html. We removed lookahead features for ISBN experiments but preserved them for experiments with the MALT parser. Analogously, we extended simple features with 3 words lookahead for the MALT parser experiments.

our estimation method and search, we perform another experiment. We use the tuned feature set and disable the feature induction abilities of the model by removing all the edges between latent variables vectors. Comparison of this restricted model with the full ISBN model shows how important the feature induction is. Also, comparison of this restricted model with the MALT parser, which uses the same set of features, indicates whether our generative estimation method and use of beam search is beneficial.

6.1 Experimental Setup

We used the CoNLL-X distributions of Danish DDT treebank (Kromann, 2003), Dutch Alpino treebank (van der Beek et al., 2002) and Slovene SDT treebank (Dzeroski et al., 2006). The choice of these treebanks was motivated by the fact that they all are freely distributed and have very different sizes of their training sets: 195,069 tokens for Dutch, 94,386 tokens for Danish and only 28,750 tokens for Slovene. As it is generally believed that discriminative models win over generative models with a large amount of training data, so we expected to see similar trend in our results. Test sets are about equal and contain about 5,000 scoring tokens.

We followed the experimental setup of the shared task and used all the information provided for the languages: gold standard part-of-speech tags and coarse part-of-speech tags, word form, word lemma (lemma information was not available for Danish) and a set of fine-grain word features. As we explained in section 3, we treated these sets of fine-grain features as an atomic value when predicting a word. However, when conditioning on words, we treated each component of this composite feature individually, as it proved to be useful on the development set. We used frequency cutoffs: we ignored any property (e.g., word form, feature or even part-of-speech tag[4]) which occurs in the training set less than 5 times. Following (Nivre et al., 2006), we used pseudo-projective transformation they proposed to cast non-projective parsing tasks as projective.

ISBN models were trained using a small development set taken out from the training set, which was used for tuning learning parameters and for early stopping. The sizes of the development sets were: 4,988 tokens for larger Dutch corpus, 2,504 tokens for Danish and 2,033 tokens for Slovene. The MALT parser was trained always using the entire training set. We expect that the mean field approximation should demonstrate better results than feed-forward approximation on this task as it is theoretically expected and confirmed on the constituent parsing task (Titov and Henderson, 2007). However, the sizes of testing sets would not allow us to perform any conclusive analysis, so we decided not to perform these comparisons here. Instead we used the mean field approximation for the smaller two corpora and used the feed-forward approximation for the larger one. Training the mean field approximations on the larger Dutch treebank is feasible, but would significantly reduce the possibilities for tuning the learning parameters on the development set and, thus, would increase the randomness of model comparisons.

All model selection was performed on the development set and a single model of each type was applied to the testing set. We used a state variable vector consisting of 80 binary variables, as it proved sufficient on the preliminary experiments. For the MALT parser we replicated the parameters from (Nivre et al., 2006) as described in detail on their web site.

The labeled attachment scores for the ISBN with tuned features (TF) and local features (LF) and ISBN with tuned features and no edges connecting latent variable vectors (TF-NA) are presented in table 1, along with results for the MALT parser both with tuned and local feature, the MST parser (McDonald et al., 2006), and the average score (Aver) across all systems in the CoNLL-X shared task. The MST parser is included because it demonstrated the best overall result in the task, non significantly outperforming the MALT parser, which, in turn, achieved the second best overall result. The labeled attachment score is computed using the same method as in the CoNLL-X shared task, i.e. ignoring punctuation. Note, that though we tried to completely replicate training of the MALT parser with the tuned features, we obtained slightly different results. The original published results for the MALT parser with tuned features were 84.8% for Danish, 78.6% for Dutch and 70.3% for Slovene. The im-

[4]Part-of-speech tags for multi-word units in the Danish treebank were formed as concatenation of tags of the words, which led to quite sparse set of part-of-speech tags.

		Danish	Dutch	Slovene
ISBN	TF	85.0	79.6	72.9
	LF	84.5	79.5	72.4
	TF-NA	83.5	76.4	71.7
MALT	TF	85.1	78.2	70.5
	LF	79.8	74.5	66.8
MST		84.8	79.2	73.4
Aver		78.3	70.7	65.2

Table 1: Labeled attachment score on the testing sets of Danish, Dutch and Slovene treebanks.

provement of the ISBN models (TF and LF) over the MALT parser is statistically significant for Dutch and Slovene. Differences between their results on Danish are not statistically significant.

6.2 Discussion of Results

The ISBN with tuned features (TF) achieved significantly better accuracy than the MALT parser on 2 languages (Dutch and Slovene), and demonstrated essentially the same accuracy on Danish. The results of the ISBN are among the two top published results on all three languages, including the best published results on Dutch. All three models, MST, MALT and ISBN, demonstrate much better results than the average result in the CoNLL-X shared task. These results suggest that our generative model is quite competitive with respect to the best models, which are both discriminative.[5] We would expect further improvement of ISBN results if we applied discriminative retraining (Henderson, 2004) or reranking with data-defined kernels (Henderson and Titov, 2005), even without introduction of any additional features.

We can see that the ISBN parser achieves about the same results with local features (LF). Local features by themselves are definitely not sufficient for the construction of accurate models, as seen from the results of the MALT parser with local features (and look-ahead). This result demonstrates that ISBNs are a powerful model for feature induction.

The results of the ISBN without edges connecting latent state vectors is slightly surprising and suggest that without feature induction the ISBN is significantly worse than the best models. This shows that the improvement is coming mostly from the ability of the ISBN to induce complex features and not from either using beam search or from the estimation procedure. It might also suggest that generative models are probably worse for the dependency parsing task than discriminative approaches (at least for larger datasets). This motivates further research into methods which combine powerful feature induction properties with the advantage of discriminative training. Although discriminative reranking of the generative model is likely to help, the derivation of fully discriminative feature induction methods is certainly more challenging.

[5]Note that the development set accuracy predicted correctly the testing set ranking of ISBN TF, LF and TF-NA models on each of the datasets, so it is fair to compare the best ISBN result among the three with other parsers.

		to root	1	2	3 - 6	> 6
Da	ISBN	95.1	95.7	90.1	84.1	74.7
	MALT	95.4	96.0	90.8	84.0	71.6
Du	ISBN	79.8	92.4	86.2	81.4	71.1
	MALT	73.1	91.9	85.0	76.2	64.3
Sl	ISBN	76.1	92.5	85.6	79.6	54.3
	MALT	59.9	92.1	85.0	78.4	47.1
Av	ISBN	83.6	93.5	87.3	81.7	66.7
	MALT	76.2	93.3	87.0	79.5	61.0
	Improv	**7.5**	**0.2**	**0.4**	**2.2**	**5.7**

Table 2: F_1 score of labeled attachment as a function of dependency length on the testing sets of Danish, Dutch and Slovene.

In order to better understand differences in performance between ISBN and MALT, we analyzed how relation accuracy changes with the length of the head-dependent relation. The harmonic mean between precision and recall of labeled attachment, F_1 measure, for the ISBN and MALT parsers with tuned features is presented in table 2. F_1 score is computed for four different ranges of lengths and for attachments directly to root. Along with the results for each of the languages, the table includes their mean (Av) and the absolute improvement of the ISBN model over MALT (Improv). It is easy to see that accuracy of both models is generally similar for small distances (1 and 2), but as the distance grows the ISBN parser starts to significantly outperform MALT, achieving 5.7% average improvement on dependencies longer than 6 word tokens. When the MALT parser does not manage to recover a long dependency, the highest scoring action it can choose is to reduce the dependent from the stack without specifying its head, thereby attaching the dependent to the root by default. This explains the relatively low F_1 scores for attachments to root (evident for Dutch and Slovene): though recall of attachment to root is comparable to that of the ISBN parser (82.4% for MALT against 84.2% for ISBN, on average over 3 languages), precision for the MALT parser is much worse (71.5% for MALT against 83.1% for ISBN, on average).

The considerably worse accuracy of the MALT parser on longer dependencies might be explained both by use of a non-greedy search method in the ISBN and the ability of ISBNs to induce history features. To capture a long dependency, the MALT parser should keep a word on the stack during a long sequence of decision. If at any point during the intermediate steps this choice seems not to be locally optimal, then the MALT parser will choose the alternative and lose the possibility of the long dependency.[6] By using a beam search, the ISBN parser can maintain the possibility of the long dependency in its beam even when other alternatives seem locally preferable. Also, long dependences are often more difficult, and may be systematically different from local dependencies. The designer of a MALT parser needs to discover predictive features for long dependencies by hand, whereas the ISBN model can automatically discover them. Thus we expect that the feature induction abilities of ISBNs have a strong effect on the accuracy of long dependences. This prediction is confirmed by the differences between the results of the normal ISBN (TF) and the restricted ISBN (TF-NA) model. The TF-NA model, like the MALT parser, is biased toward attachment to root; it attaches to root 12.0% more words on average than the normal ISBN, without any improvement of recall and with a great loss of precision. The F_1 score on long dependences for the TF-NA model is also negatively effected in the same way as for the MALT parser. This confirms that the ability of the ISBN model to induce features is a major factor in improving accuracy of long dependencies.

[6]The MALT parser is trained to keep the word as long as possible: if both **Shift** and **Reduce** decisions are possible during training, it always prefers to shift. Though this strategy should generally reduce the described problem, it is evident from the low precision score for attachment to root, that it can not completely eliminate it.

7 Related Work

There has not been much previous work on latent variable models for dependency parsing. Dependency parsing with Dynamic Bayesian Networks was considered in (Peshkin and Savova, 2005), with limited success. Roughly, the model considered the whole sentence at a time, with the DBN being used to decide which words correspond to leaves of the tree. The chosen words are then removed from the sentence and the model is recursively applied to the reduced sentence. Recently several latent variable models for constituent parsing have been proposed (Koo and Collins, 2005; Matsuzaki et al., 2005; Prescher, 2005; Riezler et al., 2002). In (Matsuzaki et al., 2005) non-terminals in a standard PCFG model are augmented with latent variables. A similar model of (Prescher, 2005) uses a head-driven PCFG with latent heads, thus restricting the flexibility of the latent-variable model by using explicit linguistic constraints. While the model of (Matsuzaki et al., 2005) significantly outperforms the constrained model of (Prescher, 2005), they both are well below the state-of-the-art in constituent parsing. In (Koo and Collins, 2005), an undirected graphical model for constituent parse reranking uses dependency relations to define the edges. Thus, it should be easy to apply a similar method to reranking dependency trees.

Undirected graphical models, in particular Conditional Random Fields, are the standard tools for shallow parsing (Sha and Pereira, 2003). However, shallow parsing is effectively a sequence labeling problem and therefore differs significantly from full parsing. As discussed in (Titov and Henderson, 2007), undirected graphical models do not seem to be suitable for history-based parsing models.

Sigmoid Belief Networks (SBNs) were used originally for character recognition tasks, but later a dynamic modification of this model was applied to the reinforcement learning task (Sallans, 2002). However, their graphical model, approximation method, and learning method differ significantly from those of this paper. The extension of dynamic SBNs with incrementally specified model structure (i.e. Incremental Sigmoid Belief Networks, used in this paper) was proposed and applied to constituent parsing in (Titov and Henderson, 2007).

8 Conclusions

We proposed a latent variable dependency parsing model based on Incremental Sigmoid Belief Networks. Unlike state-of-the-art dependency parsers, it uses a generative history-based model. We demonstrated that it achieves state-of-the-art results on a selection of languages from the CoNLL-X shared task. The parser uses a vector of latent variables to represent an intermediate state and uses relations defined on the output structure to construct the edges between latent state vectors. These properties make it a powerful feature induction method for dependency parsing, and it achieves competitive results even with very simple explicit features. The ISBN model is especially accurate at modeling long dependences, achieving average improvement of 5.7% over the state-of-the-art baseline on dependences longer than 6 words. Empirical evaluation demonstrates that competitive results are achieved mostly because of the ability of the model to induce complex features and not because of the use of a generative probability model or a specific search method. As with other generative models, it can be further improved by the application of discriminative reranking techniques. Discriminative methods are likely to allow it to significantly improve over the current state-of-the-art in dependency parsing.[7]

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. We thank Joakim Nivre and Sandra Kübler for an excellent tutorial on dependency parsing given at COLING-ACL 2006.

References

Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. 1986. *Compilers: Principles, Techniques and Tools*. Addison Wesley.

Leon Bottou. 1991. *Une approche théoretique de l'apprentissage connexionniste: Applications à la reconnaissance de la parole*. Ph.D. thesis, Université de Paris XI, Paris, France.

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

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.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proc. 43rd Meeting of Association for Computational Linguistics*, pages 173–180, Ann Arbor, MI.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proc. 1st Meeting of North American Chapter of Association for Computational Linguistics*, pages 132–139, Seattle, Washington.

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

Michael Collins. 2000. Discriminative reranking for natural language parsing. In *Proc. 17th Int. Conf. on Machine Learning*, pages 175–182, Stanford, CA.

S. Dzeroski, T. Erjavec, N. Ledinek, P. Pajas, Z. Zabokrtsky, and A. Zele. 2006. Towards a Slovene dependency treebank. In *Proc. Int. Conf. on Language Resources and Evaluation (LREC)*, Genoa, Italy.

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.

M. I. Jordan, Z.Ghahramani, T. S. Jaakkola, and L. K. Saul. 1999. An introduction to variational methods for graphical models. In Michael I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, MA.

Terry Koo and Michael Collins. 2005. Hidden-variable models for discriminative reranking. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada.

Matthias T. Kromann. 2003. The Danish dependency treebank and the underlying linguistic theory. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT)*, Vaxjo, Sweden.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Probabilistic CFG with latent annotations. In *Proceedings of the 43rd Annual Meeting of the ACL*, Ann Arbor, MI.

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.

Kevin P. Murphy. 2002. *Dynamic Belief Networks: Representation, Inference and Learning*. Ph.D. thesis, University of California, Berkeley, CA.

Radford Neal. 1992. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113.

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.

Leon Peshkin and Virginia Savova. 2005. Dependency parsing with dynamic Bayesian network. In *AAAI, 20th National Conference on Artificial Intelligence*, Pittsburgh, Pennsylvania.

Detlef Prescher. 2005. Head-driven PCFGs with latent-head statistics. In *Proc. 9th Int. Workshop on Parsing Technologies*, Vancouver, Canada.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques. In *Proc. 40th Meeting of Association for Computational Linguistics*, Philadelphia, PA.

Brian Sallans. 2002. *Reinforcement Learning for Factored Markov Decision Processes*. Ph.D. thesis, University of Toronto, Toronto, Canada.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proc. joint meeting of North American Chapter of the Association for Computational Linguistics and the Human Language Technology Conf.*, Edmonton, Canada.

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

L. van der Beek, G. Bouma, J. Daciuk, T. Gaustad, R. Malouf, G van Noord, R. Prins, and B. Villada. 2002. The Alpino dependency treebank. *Computational Linguistic in the Netherlands (CLIN)*.