It Depends on the Translation: Unsupervised Dependency Parsing via Word Alignment

Samuel Brody
Dept. of Biomedical Informatics
Columbia University
samuel.brody@dbmi.columbia.edu

Abstract

We reveal a previously unnoticed connection between dependency parsing and statistical machine translation (SMT), by formulating the dependency parsing task as a problem of word alignment. Furthermore, we show that two well known models for these respective tasks (DMV and the IBM models) share common modeling assumptions. This motivates us to develop an alignment-based framework for unsupervised dependency parsing. The framework (which will be made publicly available) is flexible, modular and easy to extend. Using this framework, we implement several algorithms based on the IBM alignment models, which prove surprisingly effective on the dependency parsing task, and demonstrate the potential of the alignment-based approach.

1 Introduction

Both statistical machine translation (SMT) and unsupervised dependency parsing have seen a surge of interest in recent years, as the need for large scale data processing has increased. The problems addressed by each of the fields seem quite different at first glance. However, in this paper, we reveal a strong connection between them and show that the problem of dependency parsing can be formulated as one of word alignment within the sentence. Furthermore, we show that the two models that are arguably the most influential in their respective fields, the IBM models 1-3 (Brown et al., 1993) and Klein and Manning's (2004) *Dependency Model with Valence* (DMV), share a common set of modeling assumptions.

Based on this connection, we develop a framework which uses an alignment-based approach for unsupervised dependency parsing. The framework is flexible and modular, and allows us to explore different modeling assumptions. We demonstrate these properties and the merit of the alignment-based parsing approach by implementing several dependency parsing algorithms based on the IBM alignment models and evaluating their performance on the task. Although the algorithms are not competitive with state-of-the-art systems, they outperform the right-branching baseline and approach the performance of DMV. This is especially surprising when we consider that the IBM models were not originally designed for the task. These results are encouraging and indicate that the alignment-based approach could serve as the basis for competitive dependency parsing systems, much as DMV did.

This paper offers two main contributions. First, by revealing the connection between the two tasks, we introduce a new approach to dependency parsing, and open the way for use of SMT alignment resources and tools for parsing. Our experiments with the IBM models demonstrate the potential of this approach and provide a strong motivation for further development. The second contribution is a publicly-available framework for exploring new alignment models. The framework uses Gibbs sampling techniques and includes our sampling-based implementations of the IBM models (see Section 3.4). The sampling approach makes it easy to modify the existing models and add new ones. The framework can be used both for dependency parsing and for bilingual word alignment.

The rest of the paper is structured as follows. In Section 2 we present a brief overview of those works in the fields of dependency parsing and alignment for statistical machine translation which are directly

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1214–1222,
MIT, Massachusetts, USA, 9-11 October 2010.

relevant to this paper. Section 3 describes the connection between the two problems, examines the shared assumptions of the DMV and IBM models, and describes our framework and algorithms. In Section 4 we present our experiments and discuss the results. We conclude in Section 5.

2 Background and Related Work

2.1 Unsupervised Dependency Parsing

In recent years, the field of supervised parsing has advanced tremendously, to the point where highly accurate parsers are available for many languages. However, supervised methods require the manual annotation of training data with parse trees, a process which is expensive and time consuming. Therefore, for domains and languages with minimal resources, unsupervised parsing is of great importance.

Early work in the field focused on models that made use primarily of the co-occurrence information of the head and its argument (Yuret, 1998; Paskin, 2001). The introduction of DMV by Klein and Manning (2004) represented a shift in the direction of research in the field. DMV is based on a linguistically motivated generative model, which follows common practice in *supervised* parsing and takes into consideration the distance between head and argument, as well as the valence (the capacity of a head word to attach arguments). Klein and Manning (2004) also shifted from a lexical representation of the sentences to representing them as part-of-speech sequences. DMV strongly outperformed previous models and was the first unsupervised dependency induction system to achieve accuracy above the right-branching baseline. Much subsequent work in the field has focused on modifications and extensions of DMV, and it is the basis for today's state-of-the-art systems (Cohen and Smith, 2009; Headden III et al., 2009).

2.2 Alignment for SMT

SMT treats translation as a machine learning problem. It attempts to learn a translation model from a parallel corpus composed of sentences and their translations. The IBM models (Brown et al., 1993) represent the first generation of word-based SMT models, and serve as a starting point for most current SMT systems (e.g., Moses, Koehn et al. 2007; Hiero, Chiang 2005). The models employ the notion of *alignment* between individual words in the source and translation. An example of such an alignment is given in Figure 1.

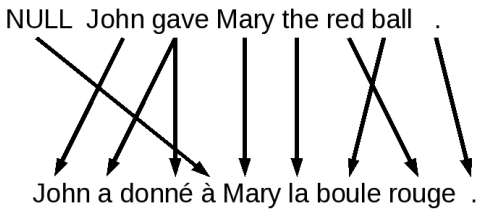

Figure 1: An example of an alignment between an English sentence (top) and its French translation (bottom).

The IBM models all seek to maximize $Pr(f|e)$, the probability of a French translation f of an English sentence e. This probability is broken down by taking into account all possible alignments a between e and f, and their probabilities:

$$Pr(f|e) = \sum_a Pr(f,a|e) \quad (1)$$

Each of the IBM models is based on the previous one in the series, and adds another level of latent parameters which take into account a specific characteristic of the data.

3 Alignment-based Dependency Parsing

3.1 The Connection

The task of dependency parsing requires finding a parse tree for a sentence, where two words are connected by an edge if they participate in a syntactic dependency relation. When dealing with unlabeled dependencies, the exact nature of the relationship is not determined. An example of a dependency parse of a sentence is given in Figure 2 (left).

Another possible formulation of the problem is as follows. Find a set of pairwise relations (s_i, s_j) connecting a dependent word s_j with its head word s_i in the sentence. This alternate formulation allows us to view the problem as one of alignment of a sentence to itself, as shown in Figure 2 (right).

Given this perspective on the problem, it makes sense to examine existing alignment models, compare them to dependency parsing models, and see if they can be successfully employed for the dependency parsing task.

Figure 2: **Left:** An example of an unlabeled dependency parse of a sentence. **Right:** The same parse, in the form of an alignment between a head words (top) and their dependents (bottom).

3.2 Comparing IBM & DMV Assumptions

Lexical Association The core assumption of IBM Model 1 is that the lexical identities of the English and French words help determine whether they should be aligned. The same assumption is made in all the dependency models mentioned in Section 2 regarding a head and its dependent (although DMV uses word classes instead of the actual words).

Location IBM Model 2 adds the consideration of difference in location between the English and French words when considering the likelihood of alignment. One of the improvements contributing to the success of DMV was the notion of distance, which was absent from previous models (see Section 3 in Klein and Manning 2004).

Fertility IBM Model 3 adds the notion of *fertility*, or the idea that different words in the source language tend to generate different numbers of words in the target language. This corresponds to the notion of *valence*, used by Klein and Manning (2004), and the other major contributor to the success of DMV (ibid.).

Null Source The IBM models all make use of an additional "null" word in every sentence, which has special status. It is attached to words in the translation that do not correspond to a word in the source. It is treated separately when calculating distance (since it has no location) and fertility. In these characteristics, it is very similar to the "root" node, which is artificially added to parse trees and used to represent the head of words which are not dependents of any other word in the sentence.

In examining the core assumptions of the IBM models, we note that there is a strong resemblance to those of DMV. The similarity is at an abstract level since the nature of the relationship that each model attempts to detect is quite different. The IBM models look for an equivalence relationship between lexical items in two languages, whereas DMV addresses functional relationships between two elements with distinct meanings. However, both attempt to model a similar set of factors, which they posit will be important to their respective tasks[1]. This similarity motivates the work presented in the rest of the paper, i.e, exploring the use of the IBM alignment models for dependency parsing. It is important to note that the IBM models do not address many important factors relevant to the parsing task. For instance, they have no notion of a parse tree, a deficit which may lead to degenerate solutions and malformed parses. However, they serve as a good starting point for exploring the alignment approach to parsing, as well as discovering additional factors that need to be addressed under this approach.

3.3 Experimental Framework

We developed a Gibbs sampling framework for alignment-based dependency parsing[2]. The traditional approach to alignment uses Expectation Maximization (EM) to find the optimal values for the latent variables. In each iteration, it considers all possible alignments for each pair of sentences, and

[1] These abstract notions (lexical association, proximity, tendencies towards few or many relations, and allowing for unassociated items) play an important role in many relation-detection tasks (e.g., co-reference resolution, Haghighi and Klein 2010).

[2] Available for download at:
`http://people.dbmi.columbia.edu/~sab7012`

chooses the optimal one based on the current parameter estimates. The sampling method, on the other hand, only considers a small change in each step - that of re-aligning a previously aligned target word to a new source. The reason for our choice is the ease of modification of such sampling models. They allow for easy introduction of further parameters and more complex probabilistic functions, as well as Bayesian priors, all of which are likely to be helpful in development[3].

Under the sampling framework, the model provides the probability of changing the alignment $A[i]$ of a target word i from a previously aligned source word j to a new one $\hat{j}$. In all the models we consider, this probability is proportional to the ratio between the scores of the old sentence alignment A and the new one $\hat{A}$, which differs from the old only in the realignment of i to $\hat{j}$.

$$P(A[i]=j \Rightarrow A[i]=\hat{j}\,) \sim \frac{P_{model}(\hat{A})}{P_{model}(A)} \qquad (2)$$

As a starting point for our dependency parsing model, we re-implemented the first three IBM models [4] in the sampling framework.

3.4 Reformulating the IBM models

IBM Model 1 According to this model, the probability of an alignment between target word i and source word $\hat{j}$ depends only on the lexical identities of the two words w_i and $w_{\hat{j}}$ respectively. This gives us equation 3.

$$P(A[i] \Rightarrow \hat{j}\,) \sim \frac{P_{model}(\hat{A})}{P_{model}(A)} = \frac{\prod_k P(w_k, w_{A[k]})}{\prod_{k'} P(w_{k'}, w_{\hat{A}[k']})}$$
$$P(A[i] \Rightarrow \hat{j}\,) \sim \frac{P(w_i, w_{\hat{j}})}{P(w_i, w_j)} \qquad (3)$$

In our implementation we assume the alignment follows a Chinese Restaurant Process (CRP), where the probability of w_i aligning to w_j is proportional to the number of times they have been aligned in the past (the rest of the data), as follows:

$$P(w_i, w_{\hat{j}}) = \frac{\#(w_i, w_j) + \alpha_1/V}{\#(*, w_j) + \alpha_1} \qquad (4)$$

Here, $\#(w_i, w_j)$ represents the number of times the target word w_i was observed to be aligned to w_j in the rest of the data, and $*$ stands for any word, V is the size of the vocabulary, and α_1 is a hyperparameter of the CRP, which can also be viewed as a smoothing factor.

IBM Model 2 The original IBM model 2 is a distortion model that assumes that the probability of an alignment between target word i and source word $\hat{j}$ depends only on the locations of the words, i.e., the values i and $\hat{j}$, taking into account the different lengths l and m of the source and target sentences, respectively. For dependency parsing, where we align sentences to themselves, $l = m$. This gives us equation 5.

$$P(A[i] \Rightarrow \hat{j}\,) \sim \frac{P_{model}(\hat{A})}{P_{model}(A)} = \frac{P(i, \hat{j}, l)}{P(i, j, l)}$$
$$P(i, \hat{j}, l) = \frac{\#(i, \hat{j}, l) + \alpha_2/D}{\#(i, *, l) + \alpha_2} \qquad (5)$$

Again, we assume a CRP when choosing a distortion value, where D is the expected number of distance values (set to 10 in our experiments), α_2 is the CRP hyperparameter, $\#(i, j, l)$ is the number of times a target word in position i was aligned to a source word in position j in sentences of length l, and $\#(i, *, l)$ is the number of times word in position i was aligned (to any source position) in sentences of length l.

Even without the need for handling different lengths for source and target sentences, this model is complex and requires estimating a separate probability for each triplet (i, j, l). In addition, the assumption that the distance distribution depends only on the sentence length and is similar for all tokens seems unreasonable, especially when dealing with part-of-speech tokens and dependency relations. Such concerns have been mentioned in the SMT literature and were shown to be justified in our experiments (see Sec. 4). For this reason, we

[3] Preliminary experiments using the EM approach via the GIZA++ toolkit (Och and Ney, 2003) resulted in similar performance to that of the sampling method for IBM Models 1 and 2. However, we were unable to explore the use of Model 3 under that framework, since the implementation of the model was strongly coupled to other, SMT-specific, optimizations and heuristics.

[4] Our implementation, as well as some core components in our framework, are based on code kindly provided by Chris Dyer.

also implemented an alternate distance model, based loosely on Liang et al. (2006). Under the alternate model, the probability of an alignment between target word i and source word $\hat{j}$ depends on the distance between them, their order, the sentence length, and the word type of the head, according to equation 6.

$$P(i,\hat{j},l) = \frac{\#[w_i,(i\text{-}\hat{j}),l] + \alpha_3/D}{\#(w_i,*,l) + \alpha_3} \quad (6)$$

IBM Model 3 This model handles the notion of fertility (or valence). Under this model, the probability of an alignment depends on how many target words are aligned to each of the source words. Each source word type $w_{\hat{j}}$, has a distribution specifying the probability of having n aligned target words. The probability of an alignment is proportional to the product of the probabilities of the fertilities in the alignment and takes into account the special status of the null word (represented by the index $j = 0$). This probability is given in Equation 7, which is based on Equation 32 in Brown et al. (1993)[5].

$$P(A) \sim \binom{l-\phi_0}{\phi_0} p_0^{l-2\phi_0} p_1^{\phi_0} \prod_{j=1}^{l} \phi_j! \frac{\#(w_j,\phi_j) + \alpha_4/F}{\#(w_j,*) + \alpha_4} \quad (7)$$

Here, ϕ_j denotes the number of target words aligned to the j-th source word in alignment A. p_1 and p_0 sum to 1 and are used to derive the probability that there will be ϕ_0 null-aligned words in a sentence containing l words[6]. $\#(w_j,\phi_j)$ represents the number of times source word w_j was observed to have ϕ_j dependent target words, $\#(w_j,*)$ is the number of times w_j appeared in the data, F is the expected number of fertility values (5 in our experiments), and α_4 is the CRP hyperparameter.

Combining the Models The original IBM models work in an incremental fashion, with each model using the output of the previous one as a starting point and adding a new component to the probability distribution. The dependency parsing framework employs a similar approach. It uses the alignments learned by the previous model as the starting point of the next and combines the probability distributions of each component via a product model. This allows for the easy introduction of new models which consider different aspects of the alignment and complement each other.

Preventing Self-Alignment When adapting the alignment approach to dependency parsing, we view the task as that of aligning a sentence to itself. One issue we must address is preventing the degenerate solution of aligning each word to itself. For this purpose we introduce a simple model into the product which gives zero probability to alignments which contain a word aligned to itself, as in equation 8.

$$P(A[i] = \hat{j}\,) = \begin{cases} 0 & \text{if } i = \hat{j} \\ \frac{1}{l-1} & \text{otherwise} \end{cases} \quad (8)$$

4 Experiments

4.1 Data

We evaluated our model on several corpora. The first of these was the Penn. Treebank portion of the Wall Street Journal (WSJ). We used the Constituent-to-Dependency Conversion Tool[7] to convert the treebank format into CoNLL format.

We also made use of the Danish and Dutch datasets from the CoNLL 2006 shared task[8]. Since we do not make use of annotation, we can induce a dependency structure on the entire dataset provided (disregarding the division into training and testing).

Following Klein and Manning (2004), we used the gold-standard part-of-speech sequences rather than the lexical forms and evaluated on sentences containing 10 or fewer tokens after removal of punctuation.

4.2 Results

Table 1 shows the results of the IBM Models on the task of directed (unlabeled) dependency parsing. We compare to the right-branching baseline used by Klein and Manning (2004). For the WSJ10 corpus, the authors reported 43.2% accuracy for DMV and 33.6% for the baseline. Although there are small

[5]The transitional version of this equation depends on whether either the old source word (j) or the new one ($\hat{j}$) are null, and is omitted for brevity. Further details can be found in Brown et al. (1993) Section 4.4 and Equation 43.

[6]For details, see Brown et al. (1993) Equation 31.

[7]nlp.cs.lth.se/software/treebank_converter/

[8]http://nextens.uvt.nl/~conll/

Corpus	M 1	M2	M3	R-br
WSJ10	25.42	35.73	39.32	32.85
Dutch10	25.17	32.46	35.28	28.42
Danish10	23.12	25.96	41.94	16.05 *

Table 1: Percent accuracy of IBM Models 1-3 (M1-3) and the right-branching baseline (R-br) on several corpora.

PoS	attachment
NN	DET
IN	NN
NNP	NNP
DET	NN
JJ	NN

PoS	attachment
NNS	JJ
RB	VBZ
VBD	NN
VB	TO
CC	NNS

Table 2: Most likely dependency attachment for the top ten most common parts-of-speech, according to Model 1.

differences in evaluation, as evidenced by the difference between our baseline scores, IBM Models 2 and 3 outperform the baseline by a large margin and Model 3 approaches the performance of DMV. On the Dutch and Danish datasets, the trends are similar. On the latter dataset, even Model 1 outperforms the right-branching baseline. However, the Danish dataset is unusual (see Buchholz and Marsi 2006) in that the alternate adjacency baseline of left-branching (also mentioned by Klein and Manning 2004) is extremely strong and achieves 48.8% directed accuracy.

4.3 Analysis

In order to better understand what our alignment model was learning, we looked at each component element individually.

Lexical Association To explore what Model 1 was learning, we analyzed the resulting probability tables for association between tokens. Table 2 shows the most likely dependency attachment for the top ten most common parts-of-speech. The model is clearly learning meaningful connections between parts of speech (determiners and adjectives to nouns, adverbs to verbs, etc.), but there is little notion of directionality, and cycles can exist. For instance, the model learns the connection between determiner and noun, but is unsure which is the head and which the dependent. A similar connection is learned between *to* and verbs in the base form (VB). This inconsistency is, to a large extent, the result of the deficiencies of the model, stemming from the fact that the IBM models were designed for a different task and are not trying to learn a well-formed tree. However, there is a strong linguistic basis to consider the directionality of these relations difficult. There is some debate among linguists as to whether the head of a noun phrase is the noun or the determiner[9] (see Abney 1987). Each can be seen as a different kind of head element, performing a different function, similarly to the multiple types of dependency relations identified in Hudson's (1990) Word Grammar. A similar case can be made regarding the head of an infinitive phrase. The infinitive form of the verb may be considered the lexical head, determining the predicate, while *to* can be seen as the functional head, encoding inflectional features, as in Chomsky's (1981) Government & Binding model[10].

Distance Models The original IBM distortion model (Model 2), which does not differentiate between words types and looks only at positions, has an accuracy of 33.43% on the WSJ10 corpus. In addition, it tends to strongly favor left-branching attachment (57.2% of target words were attached to the word immediately to their right, 22.6% to their left, as opposed to 31% and 25.8% in the gold standard). The alternative distance model we proposed, which takes into account the identity of the head word, achieves better accuracy and is closer to the gold standard balance (43.5% right and 35.3% left).

Figure 3 shows the distribution of the location of the dependent relative to the head word (at position 0) for several common parts-of-speech. It is interesting to see that singular and plural nouns (NN, NNS) behave similarly. They both have a strong preference for local attachment and a tendency towards a left-dependent (presumably the determiner, see above Table 2). Pronouns (NNP), on the other hand, are more likely to attach to the right since they are not modified by determiners. Verbs in past (VBZ) and present (VBD, VBP) forms have similar behavior, with a flatter distribution of dependent locations, whereas the base form (VB) attaches almost exclusively to the preceding token, presumably

[9] In fact, the original DMV chose the determiner as the head (see discussion in Klein and Manning 2004, Section 3).

[10] We thank an anonymous reviewer for elucidating this point.

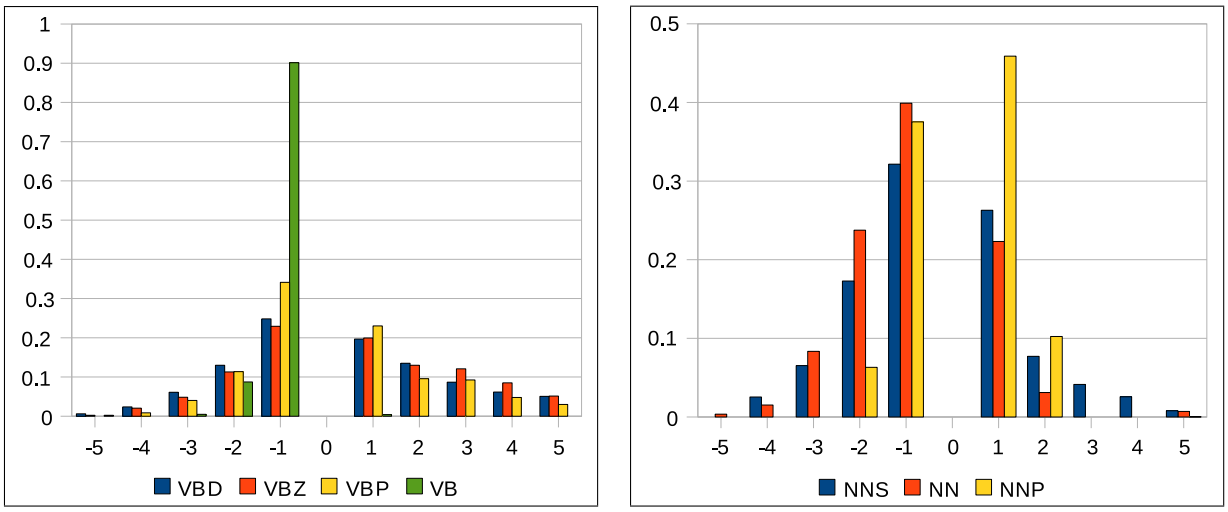

Figure 3: Distribution of head-to-dependent distance for several types of verbs (left) and nouns (right), as learned by our alternate distance model.

to (see Table 2).

Fertility Figure 4 shows the distribution of fertility values for several common parts of speech. Verbs have a relatively flat distribution with a longer tail as compared to nouns, which means they are likely to have a larger number of arguments. Once again, the base form (VB) exhibits different behavior from the other verbs forms, taking almost exclusively one argument. This is likely an effect of the strong connection between base form verbs and the preceding word *to*.

Hyper-Parameters Each of our models requires a value for its CRP hyperparameter (see Section 3.4). In this work, since parameter estimation was not our focus, we set the hyperparameters to be approximately $\frac{1}{K}$, where K is the number of possible values, according to the rule of thumb common in the literature. Specifically, we chose $\alpha_1 = 0.01, \alpha_3 = 0.05, \alpha_4 = 0.1$. We investigated the effect of these choices on performance in a separate set of experiments, which showed that small variations (up to an order of magnitude) in these parameters had little effect on the results.

In addition to the CRP parameters, Model 3 requires a value for p_1, the null fertility hyperparameter. In our experiments, we found that this hyperparameter had a very strong effect on results if it was above 0.1, creating many spurious null alignments. However, below that threshold, the effects were small. In the experiments reported here, we set $p_1 = 0.01$.

Initialization One issue with DMV, which is often mentioned, is its sensitivity to initialization. We tested our model with random initialization (uniform alignment probabilities) and with an approximation of the ad-hoc "harmonic" initialization described in Klein and Manning (2004) and found no noticeable difference in accuracy.

4.4 Discussion

The accuracy achieved by the IBM models (Table 1) is surprisingly high, given the fact that the IBM models were not designed with dependency parsing in mind. It is likely that customizing the models to the task will result in even better performance. Our findings in Section 4.3 support this hypothesis. The analysis showed that the lack of tree structure in the model impacted the learning, and therefore it is expected that a component which enforces tree structure (prevents cycles) will be beneficial.

Although it lacks an inherent notion of tree structure, the alignment-based approach has several advantages over the head-outward approach of DMV and related models. It can consider the alignment as a whole and take into account global sentence constraints, not just head-dependent relations. These may also include tree-structure constraints common to the head-outward approaches, but can be more flexible in how they are addressed. For instance,

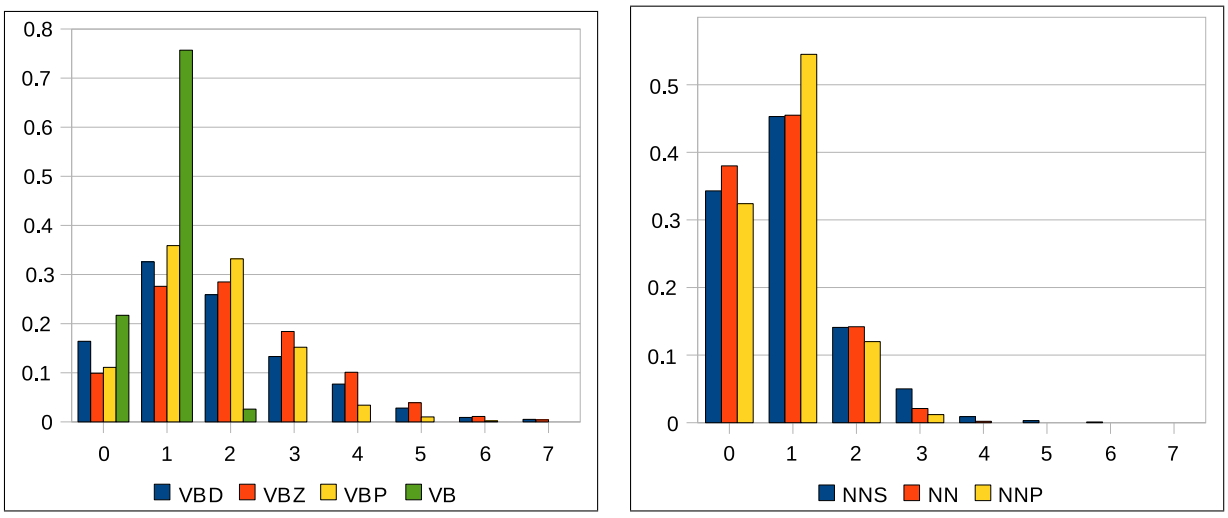

Figure 4: Distribution of fertility values for several types of verbs (left) and nouns (right), as learned by IBM Model 3.

DMV's method of modeling tree structure does not allow non-projective dependencies, whereas an alignment-based model may choose to allow or constrain non-projectivity, as learned from the data. Another advantage of our alignment-based models is the fact that they are not strongly sensitive to initialization and can be started from a set of random alignments.

5 Conclusions and Future Work

We have described an alternative formulation of dependency parsing as a problem of word alignment. This connection motivated us to explore the possibility of using alignment tools for the task of unsupervised dependency parsing. We chose to experiment with the well-known IBM alignment models which share a set of similar modeling assumptions with Klein and Manning's (2004) *Dependency Model with Valence*. Our experiments showed that the IBM models are surprisingly effective at the dependency parsing task, outperforming the right-branching baseline and approaching the accuracy of DMV. Our results demonstrate that the alignment approach can be used as a foundation for dependency parsing algorithms and motivates further research in this area.

There are many interesting avenues for further research. These include improving and extending the existing IBM models, as well as introducing new models that are specifically designed for the parsing task and represent relevant linguistic considerations (e.g., enforcing tree structure, handling crossing dependencies, learning left- or right-branching tendencies).

In Spitkovsky et al. (2010), the authors show that a gradual increase in the complexity of the data can aid the learning process. The IBM approach demonstrated the benefit of a gradual increase of *model* complexity. It would be interesting to see if the two approaches could be successfully combined.

Finally, although we use our framework for dependency parsing, the sampling approach and the framework we developed can be used to explore new models for bilingual word alignment. Furthermore, an alignment-based parsing method is expected to integrate well with SMT bi-lingual alignment models and may, therefore, be suitable for combined models which use parse trees to improve word alignment (e.g., Burkett et al. 2010).

Acknowledgments

I would like to thank Chris Dyer for providing the basis for the sampling implementation. I would also like to thank Chris, Adam Lopez, Trevor Cohn, Adam Faulkner and the anonymous reviewers for their time and effort and their helpful comments and suggestions.

References

Abney, Steven. 1987. *The English Noun Phrase in its Sentential Aspect*. Ph.D. thesis, Massachusetts Insti-

tute of Technology.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* 19(2):263–311.

Buchholz, Sabine and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *In Proc. of CoNLL*. pages 149–164.

Burkett, David, John Blitzer, and Dan Klein. 2010. Joint parsing and alignment with weakly synchronized grammars. In *North American Association for Computational Linguistics*. Los Angeles.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, pages 263–270.

Chomsky, Noam. 1981. *Lectures on government and binding : the Pisa lectures / Noam Chomsky*.

Cohen, Shay B. and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, pages 74–82.

Haghighi, Aria and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, pages 385–393.

Headden III, William P., Mark Johnson, and David McClosky. 2009. Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Boulder, Colorado, pages 101–109.

Hudson, R. 1990. *English Word Grammar*. Basil Blackwell, Oxford.

Klein, Dan and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, page 478.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Morristown, NJ, USA, pages 177–180.

Liang, Percy, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Association for Computational Linguistics, New York City, USA, pages 104–111.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1):19–51.

Paskin, Mark A. 2001. Grammatical bigrams. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*. MIT Press, pages 91–97.

Spitkovsky, Valentin I., Hiyan Alshawi, and Daniel Jurafsky. 2010. From Baby Steps to Leapfrog: How "Less is More" in unsupervised dependency parsing. In *Proc. of NAACL-HLT*.

Yuret, D. 1998. *Discovery of linguistic relations using lexical attraction*. Ph.D. thesis, Department of Computer Science and Electrical Engineering, MIT.