Supertagging with Factorial Hidden Markov Models*

Srivatsan Ramanujam[a] and Jason Baldridge[b]

[a] Department of Computer Sciences, The University of Texas at Austin,
Austin, TX 78712, USA
vatsan@cs.utexas.edu
[b] Department of Linguistics, The University of Texas at Austin,
Austin, TX 78712, USA
jbaldrid@mail.utexas.edu

Abstract. Factorial Hidden Markov Models (FHMM) support joint inference for multiple sequence prediction tasks. Here, we use them to jointly predict part-of-speech tag and supertag sequences with varying levels of supervision. We show that supervised training of FHMM models improves performance compared to standard HMMs, especially when labeled training data is scarce. Secondly, we show that an FHMM and a maximum entropy Markov model in a single step co-training setup improves the performance of both models when there is limited labeled training data. Finally, we find that FHMMs trained from tag dictionaries rather than labeled examples also perform better than a standard HMM.

Keywords: Hidden Markov Models, Bayesian Models, Categorial Grammar, Supertagging

1 Introduction

For many sequence prediction tasks in Natural Language Processing, modeling dependencies between individual predictions can be used to improve prediction accuracy of the sequence as a whole. For example, chunking involves identifying sequences of words in a sentence that are part of syntactically related non-overlapping, non-recursive phrases. An effective representation for this task involves assigning an individual part-of-speech (POS) tag and chunk tag to each word and deriving the actual chunks from these word specific labels. In these sequences, many of the POS and chunk tags are correlated, so joint inference can be quite useful.

Supertagging (Bangalore and Joshi, 1999), involves assigning lexical entries to words based on lexicalized grammatical theory such as Combinatory Categorial Grammar (CCG) (Steedman, 2000; Steedman and Baldridge, 2009). For example, the English verb *join* has the POS VB and the CCG category $((S_b\backslash NP)/PP)/NP$ in CCGbank (Hockenmaier and Steedman, 2007). This category indicates that *join* requires a noun phrase to its left, another to its right, and a prepositional phrase to the right of that. Every lexical item has as many supertags as the number of different syntactic contexts in which the item can appear, so supertags are far more detailed and numerous than POS tags. Recently there is increased interest on supertagging beyond their standard use as a pre-parsing step (Clark and Curran, 2007)—for example, they are being used as features in machine translation (Birch *et al.*, 2007; Hassan *et al.*, 2007).

Chunking and supertagging can be modeled using a two-stage cascade of Hidden Markov Models (HMMs) (Rabiner, 1989). POS tags are first predicted from the observed words in the first stage; then the chunk tags or supertags are predicted from those POS tags in the next stage. Alternatively, both sequences can be jointly predicted with Factorial Hidden Markov Models (FHMMs) (Ghahramani and Jordan, 1998), thereby preventing propagation of errors. Here, we apply

* We would like to thank Sharon Goldwater and Ray Mooney for helpful feedback and suggestions. This work was supported by the Morris Memorial Grant from the New York Community Trust.

FHMMs to supertagging for the categories defined in CCGbank for English. Fully supervised maximum entropy Markov models have been used for cascaded prediction of POS tags followed by supertags (Clark and Curran, 2007). Here, we learn supertaggers given only a POS tag dictionary and supertag dictionary or a small amount of material labeled with both types of information. Previous work has used Bayesian HMMs to learn taggers for both POS tagging (Goldwater and Griffiths, 2007) and supertagging (Baldridge, 2008) separately. Modeling them jointly has the potential to produce more robust and accurate supertaggers trained with less supervision and thereby potentially help in the creation of useful models for new languages and domains.

Our results show that joint inference improves supervised supertag prediction (compared to HMMs), especially when labeled training data is scarce. Secondly, when training data is limited, the generative FHMMs and a maximum entropy Markov model (a discriminative model like C&C) can bootstrap each other, in a single round co-training setup, to complement each other. Finally, FHMMs trained on tag dictionaries also outperform standard HMMs, thereby providing a stronger basis for learning accurate supertaggers with less supervision.

2 Data

CCG is a lexicalized grammar formalism in which the grammatical constituents have types that are detailed categories like NP, $(S\backslash NP)/NP$, and $(N\backslash N)/(S/NP)$ that specify, among other things, the sub-categorization requirements of the constituent. Every word is associated with a *lexical category*; strings of adjacent words and word sequences may then be joined via universal rules of category combination. An analysis of a sentence is complete when a single derived category spanning all words in the sentence is reached, as shown in Figure 1.

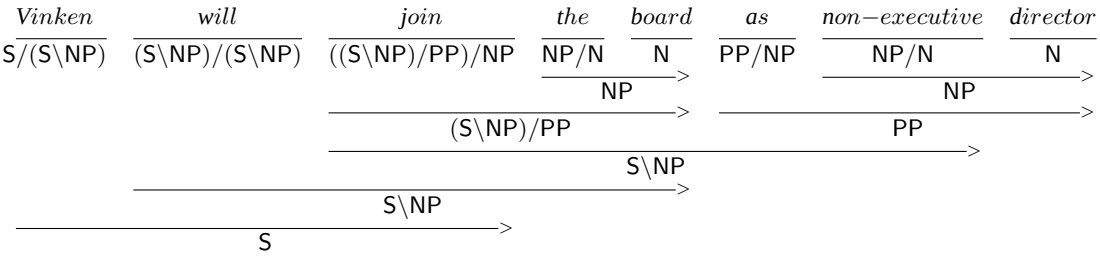


Figure 1: CCG derivation: lexical categories (the interest of this paper) appear below words and derived categories appear below lines showing combination using the forward application rule ($X/Y\ Y \rightarrow X$).

Accurately assigning lexical categories to words is the key to fast parsing for CCG. Clark and Curran (2007) use a maximum entropy Markov Model to predict lexical categories before fully parsing a sentence based on those categories. In another light, supertagging for CCG can also be seen as a way of generalizing a lexicon by identifying categories for unseen words or unobserved word/category pairs. The performance of the C&C supertagger relies on the existence of CCGbank, which itself is a semi-automated conversion of phrase structure analyses of the Penn Treebank (Marcus *et al.*, 1994) into CCG analyses. CCGbank, not to mention the original Penn Treebank, required considerable manual effort to create. It is thus of interest to build accurate supertaggers, and also to do so with as little supervision as possible in order to support the creation of grammars and annotated resources for other domains and languages.

Table 1 summarizes some attributes and the ambiguities of the POS and CCG types and tokens in CCGbank. The larger size of the CCG tag set translates to a greater per token ambiguity in the prediction of CCG tags. Supertagging is thus generally a harder problem than POS tagging.

3 The Models

We consider three different models in this paper: an HMM model and two FHMM models which we call FHMMA and FHMMB. In the standard HMM shown in Figure 2(a), a tag (POS tag or

Table 1: CCGbank statistics.

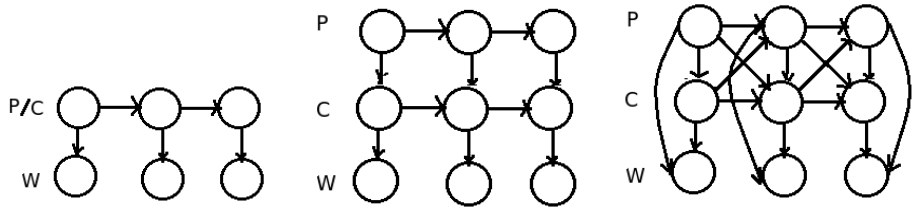
No. Training Sentences	38015
No. Testing Sentences	5435
No. Unique POS Tags	48
No. Unique CCG Tags	1241
%OOV Words (in Test)	2.86

(a) CCGbank Dataset

Avg. POS type ambiguity	1.16
Avg. POS token ambiguity	2.19
Avg. CCG type ambiguity	1.71
Avg. CCG token ambiguity	18.71
Avg. Pair token ambiguity	20.19

(b) Ambiguity in training set

CCG supertag) generates a word and the following tag at each time slice. Baldridge (2008) used such a bi-tag HMM for supertagging. In the FHMMA model (Figure 2(b)), each POS tag generates the following POS tag and the current CCG supertag; each supertag in turn generates the following supertag and the current word. The FHMMB (Figure 2(c)) has a greater interlinking of POS and CCG tags in adjacent time slices: every POS tag and supertag is dependent on both the preceding POS tag and supertag, and both POS tags and supertags jointly generate the current word.



(a) HMM (b) FHMMA (c) FHMMB

Figure 2: The Models

We use Bayesian inference for HMMs following Goldwater and Griffiths (2007) and Johnson *et al.* (2007), with symmetric Dirichlet priors for the transition and emission distributions of each of the three models. Such bitag HMMs can be formulated as:

$$\begin{aligned} t_i|t_{i-1}=t,\tau^{(t,t')} &\sim Mult(t) \\ w_i|t_i=t,\omega^{(t)} &\sim Mult(\omega^{(t)}) \\ \tau^{(t,t')}|\alpha, &\sim Dirichlet(\alpha) \\ \omega^{(t)}|\beta, &\sim Dirichlet(\beta) \end{aligned}$$

Here, t_i and w_i refer to the i'th tag and word and τ refers to the state transition distribution and ω refers to the word emission distribution.

The Forward-Backward algorithm standardly used for HMMs is intractable for FHMMMs due to their larger state space. We thus use Gibbs sampling, a Markov Chain Monte Carlo method that is commonly used for inference in Bayesian graphical models (Besag, 2004; Gao and Johnson, 2007). The Gibbs sampling equations for a POS and CCG pair in each of the models are summarized in Figure 3. For the HMM model, the POS and CCG tags are sampled independently of each other. For the FHMMs, the interlinks between the POS and CCG nodes in the graphical model, determines the interdependency during the joint inference of the POS and CCG tag sequences.

4 Supervised Supertagging Experiments

We consider two supervised training scenarios here.

(1) $P(t_i|MB(t_i)) \propto P(t_i|t_{i-1})P(t_{i+1}|t_i)P(w_i|t_i)$
(2) $P(c_i|MB(c_i)) \propto P(c_i|c_{i-1})P(c_{i+1}|c_i)P(w_i|c_i)$
(3) $P(t_i, c_i|MB(t_i, c_i)) \propto P(t_i|t_{i-1})P(c_i|t_i, c_{i-1})P(t_{i+1}|t_i)P(c_{i+1}|t_{i+1}, c_i)P(w_i|c_i)$
(4) $P(t_i, c_i|MB(t_i, c_i)) \propto P(t_i|t_{i-1}, c_{i-1})P(c_i|t_{i-1}, c_{i-1}, t_i)P(t_{i+1}|t_i, c_i)P(c_{i+1}|t_i, c_i, t_{i+1})P(w_i|t_i, c_i)$

Figure 3: Sampling equations for each model: (1) and (2) for HMM, (3) for FHHMA and (4) for FHMMB. $MB(.)$ denotes the Markov Blanket of a node in the graphical model.

4.1 Supertagging With Varying Amounts of Training Data

In this experiment, we use the training and test sets used by Baldridge (2008) from CCGbank. We vary the amount of training material by using 100, 1000, 10,000 and all 38015 training set sentences. We also vary the transition prior α choosing $\alpha = 1.0$ and $\alpha = 0.05$ on the CCG tags. The emission prior β was held constant at 1.0. The results of these experiments for $\alpha = 0.05$ are tabulated in Table 3(a). For comparison, we also show the results of the C&C supertagger of Clark and Curran (2007) in Table 3(b).

The parameter α, which determines the sparsity of the transition matrix, has been reported to have a greater influence on the performance of the tagger in Goldwater and Griffiths (2007) in weakly supervised POS tagging. We also observed this in supervised supertagging, in the models HMM and FHMMB. The HMM model and FHMMB showed a slight dip in their performance for $\alpha = 1.0$ while FHMMA did slightly better.

What stands out in these results is the performance of the FHMM models with minimal amount of training data (for 100 sentences, FHMMB is quite close to the discriminatively trained C&C supertagger). The FHMMA model achieves a 22% absolute accuracy improvement for CCG tags (ambiguous types alone) when compared to the HMM model and the FHMMB model achieves a 41% improvement compared to the HMM model.

Table 2: Supervised supertagging performance for $\alpha = 0.05$

No. Sen.	HMM	FHMMA	FHMMB
100	17.23	39.79	58.07
1000	46.02	59.93	74.31
10000	70.49	73.79	83.85
ALL	76.65	76.98	86.21

(a) Ambiguous types alone

No. Sen.	HMM	FHMMA	FHMMB	C&C
100	40.86	55.55	67.45	70.25
1000	59.04	68.39	78.01	83.25
10000	75.27	77.93	85.78	89.67
ALL	79.95	80.06	87.68	91.36

(b) Inclusive of ambiguous types.

State-of-the-art POS taggers report accuracies in the range of $96-97\%$; our model FHMMB was comparable (95.35% for $\alpha = 0.05$ and 94.41 for $\alpha = 1.0$). The FHMMA model and the HMM model achieved 91% and 92.5% accuracy on POS tags, respectively. The accuracy of our HMM is lower than the performance of Baldridge (2008) for supertags. We attribute this to better tag-specific smoothing in his model for emissions, compared to our use of a symmetric parameter for all tags. We stress that our interest here is in evaluating the advantage of joint inference over POS tags and supertags rather than direct supertag prediction while holding all other modeling considerations equal.

4.2 Single Round Co-Training Experiments

In this section, we use the FHMMB model and the C&C model in a single round of co-training. The idea behind co-training (Blum and Mitchell, 1998) is that when two models learn different *views* of the task, and are not correlated in the errors they make, they may compliment each other

to boost each other's performance. Thus, provided with a large amount of unannotated data, we could use co-training iteratively to enhance the prediction performance of the two models. For example, Clark *et al.* (2003) co-train the C&C tagger and the TNT tagger (Brants, 2000) and obtain significant performance improvements for POS tagging.

Here, we do not perform co-training, strictly speaking. Instead, we complete a *single* round in which one model is trained on a small number of sentences and then is used to label all remaining unannotated examples; the entire set is then used by the other model for training.

Table 3(a) shows the results of bootstrapping the C&C supertagger with the FHMMB model trained on 25, 50 and 100 annotated sentences respectively. For comparison, the standalone performance of C&C is also shown. The FHMMB model was used to annotate the remaining sentences in the training set and the C&C supertagger was trained on this larger annotated dataset and its prediction performance was tested on the test set. Table 3(b) shows the results of using the C&C supertagger trained on 25, 50 and 100 annotated sentences, to bootstrap the FHMMB model. Again, for comparison, the standalone performance of the FHMMB model is also shown. The C&C supertagger is used to annotate the remaining sentences in the training dataset and the FHMMB model is trained on them and its prediction performance is tested on the test dataset.

Table 3: Single round co-training with 25, 50 and 100 annotated sentences.

Num Sen.	Alone	Bootstrapped
25	53.86	60.18
50	62.03	63.97
100	70.25	69.35

(a) C&C bootstrapped with FHMMB

Num Sen.	Alone	Bootstrapped
25	58.25	56.72
50	62.36	65.37
100	67.45	71.7

(b) FHMMB bootstrapped with C&C

C&C bootstrapped by the FHMMB model outperforms C&C alone when training on the same sentences. This makes it ideal for bootstrapping more powerful discriminative models like C&C. Clearly, from Table 3(a), we see that co-training helps the C&C supertagger in improving its supertagging performance, with minimal supervision (25, 50 sentences). In Table 3(b), we see that the C&C supertagger helps in boosting the performance of the FHMMB model. These results suggest that further experiments applying standard multi-round co-training could improve both models considerably. Finally, note that FHMMB's lone performance of 58.25% with 25 seed sentences is considerably better than C&C's lone performance of 53.86% with the same seed set.

5 Weakly Supervised Supertagging

Since annotation is costly, we are interested in automatic annotation of unlabeled sentences with minimal supervision. In the weakly supervised learning setting, we are provided with a lexicon that lists possible POS tags and supertags for many, though not all, words.

We draw the initial sample of CCG tag sequences corresponding to the observation sequence, using probabilities based on *grammar informed initialization* (Baldridge, 2008). We consider the prior probability of occurrence of categories based on their *complexity*: given a lexicon L, the probability of a category c_i is inversely proportional to its complexity:

$$\Lambda_i = (1/complexity(c_i))/\sum_{j \in L}(1/complexity(c_j)) \quad (1)$$

where $complexity(c_i)$ is defined as the number of sub-categories contained in category c_i.

The POS tag corresponding to an observed word w_i is drawn uniformly at random from the set of all tags corresponding to w_i in the dictionary. For the FHHMs, we first draw a POS tag t_i

corresponding to a word w_i uniformly at random from the tag dictionary of w_i and then from the set of all CCG tags that have occurred with t_i and w_i in the dictionary, we randomly sample a CCG tag c_i based on its complexity, as defined above.

5.1 Effect of Frequency Cut-off on Supertags

Any category c, that occurs less than $k\%$ of the times with a word type w, is removed from the tag dictionary of that word, when the lexicon is constructed. This is in fact a form of supervision, which we use here as an *oracle* to explore the effect of reducing lexical ambiguity.

Results of this experiment for $\alpha = 1.0$, on ambiguous CCG categories, are tabulated in Table 5(a). The results for $\alpha = 0.05$ is shown in Table 6(a). We also report the CCG accuracy values inclusive of unambiguous types in Table 5(b) for $\alpha = 1.0$ and Table 6(b) respectively.

The performance of the HMM model (31%) in Table 5(a) without any frequency cut-off on the CCG categories, is comparable to the bitag HMM of Baldridge (2008) that uses variational Bayes EM (33%). Our complexity based initialization is not directly comparable to the results in Baldridge (2008) because the values there are based on a weighted combination of complexity based initialization and modified transition priors based on the CCG formalism. However, it is encouraging to see that when there is no cut-off based filtering of the categories, FHMMB (47.98%) greatly outperforms the HMM-EM model of Baldridge (2008). It is however, quite short of the 56.1% accuracy achieved by the model of Baldridge (2008) that uses grammar informed initialization (combination of category based initialization along with category transition rules).

Table 4: Weakly Supervised Supertagging with $\alpha = 1.0$.

CCG cut-off	HMM	FHMMA	FHMMB
0.1	63.99	47.16	60.66
0.01	65.77	45.72	61.61
0.001	46.49	39.51	52.60
None	30.89	37.36	47.98

(a) Ambiguous types

CCG cut-off	HMM	FHMMA	FHMMB
0.1	78.78	70.23	76.99
0.01	74.85	60.86	71.95
0.001	55.70	50.07	60.62
None	37.46	46.93	52.95

(b) Inclusive of unambiguous types

Table 5: Weakly Supervised Supertagging with $\alpha = 0.05$.

CCG cut-off	HMM	FHMMA	FHMMB
0.1	59.21	47.15	62.38
0.01	45.42	42.73	51.08
0.001	27.41	36.03	35.94
None	23.02	30.5	34.03

(a) Ambiguous types

CCG cut-off	HMM	FHMMA	FHMMB
0.1	76.44	70.36	77.96
0.01	60.68	58.91	64.63
0.001	40.31	47.3	47.19
None	31.74	41.3	40.67

(b) Inclusive of unambiguous types

Without any frequency cut-off on CCG categories, FHMMB achieves over 17% improvement in the prediction accuracy of ambiguous CCG categories, in comparison with the HMM. The HMM performs much better when there is a high level of frequency based filtering of the categories. However, recall that frequency based filtering of categories is a strong form of supervision that we use here only as an oracle and which one could not expect to have in real world tag dictionaries. The POS accuracies in these experiments were 83.5-85%, 84.5-86.2% and 78.3-78.4% for models FHMMB, FHMMA and HMM respectively (without any frequency cut-off).

In the weakly supervised setting, the choice of the transition prior α of 0.05 lead to severe degradation in the prediction accuracy of CCG tags. Unlike POS tagging, where a symmetric transition prior of $\alpha = 0.05$ captured the sparsity of the tag transition distribution (Goldwater and Griffiths, 2007), in supertagging the transition priors are asymmetric. We expect that CCG transition rules (Baldridge, 2008) when encoded as category specific transition priors, will lead to better performance with the FHMMs.

6 Related Work

This paper follows the work of Duh (2005), Baldridge (2008) and Goldwater and Griffiths (2007). Duh (2005) uses FHMMs for jointly labeling the POS and NP chunk tags for the CoNLL2000 dataset (Sang *et al.*, 2000). His is a fully supervised model for a simpler task. We address the harder problem of supertagging in this paper and especially in the weakly supervised setting, with FHMMs.

Goldwater and Griffiths (2007) uses a Bayesian tritag HMM (BHMM) for POS tagging and considers three different scenarios: (1) a weakly supervised setting with fixed hyperparameters α and β, (2) hyper parameter inference (learning the optimal values for α and β) and (3) hyper parameter inference with varying corpus size and dictionary knowledge. Our bitag HMM achieved results close to what was reported by her BHMM on a random 24000 word subset of the WSJ. In all our experiments, we have kept the test set separate from the training set from which the dictionary was built; this distinction is not made by Goldwater. Again, our work focuses on the harder problem of supertagging.

McCallum *et al.* (2003) have also used a factorial model for performing joint labeling of the POS and chunk tags but by using Dynamic Conditional Random Fields (DCRF). The advantage of using an FHMM over DCRF is the the ability to use less supervision in training the model. Even in the supervised training scenario, FHMM has the advantage of lower training time when compared to discriminative training models like DCRF.

7 Conclusion

We demonstrated that joint inference in supertagging, boosts the prediction accuracy of both POS and CCG tags by a considerable margin. The improvement is more significant when training data is scarce. The results from the single round co-training experiments were encouraging. The generative FHMM model is able to rival a discriminative model like the C&C supertagger, when more labeled sentences are made available by a bootstrapped supertagger.

To the best of our knowledge, this is the first work on joint inference in the Bayesian framework for supertagging. There is plenty of scope for further improvements. Overall, the discriminative C&C supertagger outperforms the FHMMs in all supervised settings. Despite this, the FHMMs are suited for estimating models with less supervision, such as from tag dictionaries alone and incorporating more informative prior distributions such as those in Baldridge (2008). This may make them more appropriate for developing CCGbanks for other languages and domains. Furthermore, Bayesian inference is modular and extensible, so our models could be supplemented by finding optimal values of the hyperparameters α (for POS tags) and β.

References

Baldridge, J. 2008. Weakly Supervised Supertagging with Grammar Informed Initialization. Proc. of COLING-2008. Manchester, UK.

Bangalore, S. and A. K. Joshi. 1999. Supertagging: An Approach to Almost Parsing. Computational Linguistics 25(2): 237-265.

Besag, J. 2004. Introduction to Markov Chain Monte-Carlo Methods. Mathematical Foundations of Speech and Language Processing. Springer, New York. 247-270.

Birch, A., M. Osborne and P. Koehn. 2007. CCG Supertags in Factored Statistical Machine Translation. Proc. of 2nd Workshop on Statistical Machine Translation.

Blum, A. and T. Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. Proc. of the 1998 Conference on Computational Learning Theory, 92-100.

Brants, T. 2000. TNT - a Statistical Part-of-Speech Tagger. Proc. of the 6th Conference on Applied Natural Language Processing, 224-231.

Clark, S. and J. Curran. 2007. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. Computational Linguistics, 33(4).

Clark, S., J. Curran and M. Osborne. 2003. Bootstrapping POS taggers using Unlabeled Data. Proceedings of CoNLL-2003, 49-55.

Duh, K. 2005. Joint Labelling of Multiple Sequences: A Factorial HMM Approach. 43rd Annual Meeting of the Assoc. for Computational Linguistics (ACL 2005), Student Research Workshop, Michigan..

Gao, J. and M. Johnson. 2008. Comparison of Bayesian Estimators for unsupervised Hidden Markov Model POS Taggers. Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 344-352.

Ghahramani, Z. and M. Jordan. 1998. Factorial Hidden Markov Models. Machine Learning, 29(2-3): 245-273.

Goldwater, S. and T. Griffiths. 2007. A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging. Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07).

Hassan, H., K. Sima'an and A. Way. 2007. Supertagged Phrase-Based Statistical Machine Translation. Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)

Hockenmaier, J. and M. Steedman. 2007. CCG-bank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. Computational Linguistics, 33(3):355-396.

Johnson, M., T. Griffiths and S. Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics. 139-146.

Johnson, M. 2007. Why doesn't EM Find Good HMM POS-Taggers ? Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 296-305.

Marcus, M. P., B. Santorini and M. A. Marcinkiewicz. 1994. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, 19.

Mccallum, A., K. Rohanimanesh and C. Sutton. 2003. Dynamic Conditional Random Fields for Jointly Labeling Multiple Sequences. NIPS Workshop on Syntax, Semantics and Statistics.

Rabiner, L. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 77, 257-286.

Sang, E. T. K. and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 Shared Task: Chunking. Proc. of CoNLL-2000 and LLL-2000, Lisbon, Portugal.

Steedman, M. 2000. The Syntactic Process. The MIT Press, Cambridge Mass.

Steedman, M. and J. Baldridge. To appear. Combinatory Categorial Grammar. Robert Borsley and Kersti Borjars (eds.) Constraint-based approaches to grammar: alternatives to transformational syntax. Oxford: Blackwell.