Evaluating Unsupervised Part-of-Speech Tagging for Grammar Induction

William P. Headden III, David McClosky, Eugene Charniak
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University
Providence, RI 02912
{headdenw,dmcc,ec}@cs.brown.edu

Abstract

This paper explores the relationship between various measures of unsupervised part-of-speech tag induction and the performance of both supervised and unsupervised parsing models trained on induced tags. We find that no standard tagging metrics correlate well with unsupervised parsing performance, and several metrics grounded in information theory have no strong relationship with even supervised parsing performance.

1 Introduction

There has been a great deal of recent interest in the unsupervised discovery of syntactic structure from text, both parts-of-speech (Johnson, 2007; Goldwater and Griffiths, 2007; Biemann, 2006; Dasgupta and Ng, 2007) and deeper grammatical structure like constituency and dependency trees (Klein and Manning, 2004; Smith, 2006; Bod, 2006; Seginer, 2007; Van Zaanen, 2001). While some grammar induction systems operate on raw text, many of the most successful ones presume prior part-of-speech tagging. Meanwhile, most recent work in part-of-speech induction focuses on increasing the degree to which their tags match hand-annotated ones such as those in the Penn Treebank.

In this work our goal is to evaluate how improvements in part-of-speech tag induction affects grammar induction. Using several different unsupervised taggers, we induce tags and train three grammar induction systems on the results.

We then explore the relationship between the performance on common unsupervised tagging metrics and the performance of resulting grammar induction systems. Disconcertingly we find that they bear little to no relationship.

This paper is organized as follows. In Section 2 we discuss unsupervised part-of-speech induction systems and common methods of evaluation. In Section 3, we describe grammar induction in general and discuss the systems with which we evaluate taggings. We present our experiments in Section 4, and finally conclude in Section 5.

2 Part-of-speech Tag Induction

Part-of-speech tag induction can be thought of as a clustering problem where, given a corpus of words, we aim to group word tokens into syntactic classes.

Two tasks are commonly labeled unsupervised part-of-speech induction. In the first, tag induction systems are allowed the use of a tagging dictionary, which specifies for each word a set of possible parts-of-speech (Merialdo, 1994; Smith and Eisner, 2005; Goldwater and Griffiths, 2007). In the second, only the word tokens and sentence boundaries are given. In this work we focus on this latter task to explore grammar induction in a maximally unsupervised context.

Tag induction systems typically focus on two sorts of features: distributional and morphological. Distributional refers to what sorts of words appear in close proximity to the word in question, while morphological refers to modeling the internal structure of a word. All the systems below make use of distributional information, whereas only two use morphological features.

We primarily focus on the metrics used to evaluate induced taggings. The catalogue of recent part-of-speech systems is large, and we can only test

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 329–336
Manchester, August 2008

the tagging metrics using a few systems. Recent work that we do not explore explicitly includes (Biemann, 2006; Dasgupta and Ng, 2007; Freitag, 2004; Smith and Eisner, 2005). We have selected a few systems, described below, that represent a broad range of features and techniques to make our evaluation of the metrics as broad as possible.

2.1 Clustering using SVD and K-means

Schütze (1995) presents a series of part-of-speech inducers based on distributional clustering. We implement the baseline system, which Klein and Manning (2002) use for their grammar induction experiments with induced part-of-speech tags. For each word type w in the vocabulary V, the system forms a feature row vector consisting of the number of times each of the F most frequent words occur to the left of w and to the right of w. It normalizes these row vectors and assembles them into a $|V| \times 2F$ matrix. It then performs a Singular Value Decomposition on the matrix and rank reduces it to decrease its dimensionality to d principle components ($d < 2F$). This results in a representation of each word as a point in a d dimensional space. We follow Klein and Manning (2002) in using K-means to cluster the d dimensional word vectors into parts-of-speech. We use the $F = 500$ most frequent words as left and right context features, and reduce to a dimensionality of $d = 50$. We refer to this system as SVD in our experiments.

The other systems described in Schütze (1995) make use of more complicated feature models. We chose the baseline system primarily to match previous evaluations of grammar induction using induced tags (Klein and Manning, 2002).

2.2 Hidden Markov Models

One simple family of models for part-of-speech induction are the Hidden Markov Models (HMMs), in which there is a sequence of hidden state variables $t_1 ... t_n$ (for us, the part-of-speech tags). Each state t_i is conditioned on the previous $n-1$ states $t_{i-1} ... t_{i-n+1}$, and every t_i emits an observed word w_i conditioned on t_i. There is a single start state that emits nothing, as well as a single stop state, which emits an end-of-sentence marker with probability 1 and does not transition further. In our experiments we use the bitag HMM, in which each state t_i depends only on state t_{i-1}.

The classic method of training HMMs for part-of-speech induction is the Baum-Welch (Baum, 1972) variant of the Expectation-Maximization (EM) algorithm, which searches for a local maximum in the likelihood of the observed words.

Other methods approach the problem from a Bayesian perspective. These methods place Dirichlet priors over the parameters of each transition and emission multinomial. For an HMM with a set of states T and a set of output symbols V:

$$\forall t \in T \quad \theta_t \sim Dir(\alpha_1, ...\alpha_{|T|}) \quad (1)$$

$$\forall t \in T \quad \eta_t \sim Dir(\beta_1, ...\beta_{|V|}) \quad (2)$$

$$t_i | t_{i-1},\ \theta_{t_{i-1}} \sim Multi(\theta_{t_{i-1}}) \quad (3)$$

$$w_i | t_i,\ \eta_{t_i} \sim Multi(\eta_{t_i}) \quad (4)$$

One advantage of the Bayesian approach is that the prior allows us to bias learning toward sparser structures, by setting the Dirichlet hyperparameters α, β to a value less than one (Johnson, 2007; Goldwater and Griffiths, 2007). This increases the probability of multinomial distributions which put most of their mass on a few events, instead of distributing them broadly across many events. There is evidence that this leads to better performance on some part-of-speech induction metrics (Johnson, 2007; Goldwater and Griffiths, 2007).

There are both MCMC and variational approaches to estimating HMMs with sparse Dirichlet priors; we chose the latter (Variational Bayes or VB) due to its simple implementation as a minor modification to Baum-Welch. Johnson (2007) evaluates both estimation techniques on the Bayesian bitag model; Goldwater and Griffiths (2007) emphasize the advantage in the MCMC approach of integrating out the HMM parameters in a tritag model, yielding a tagging supported by many different parameter settings.

Following the setup in Johnson (2007), we initialize the transition and emission distributions to be uniform with a small amount of noise, and run EM and VB for 1000 iterations. We label these systems as HMM-EM and HMM-VB respectively in our experiments. In our VB experiments we set $\alpha_i = \beta_j = 0.1, \forall i \in \{1, ..., |T|\}, j \in \{1, ..., |V|\}$, which yielded the best performance on most reported metrics in Johnson (2007). We use maximum marginal decoding, which Johnson (2007) reports performs better than Viterbi decoding.

2.3 Systems with Morphology

Clark (2003) presents several part-of-speech induction systems which incorporate morphological as well as distributional information. We use the

implementation found on his website.[1]

2.3.1 Ney-Essen with Morphology

The simplest model is based on work by (Ney et al., 1994). It uses a bitag HMM, with the restriction that each word type in the vocabulary can only be generated by a single part-of-speech. Thus the tag induction task here reduces to finding a multiway partition of the vocabulary. The learning algorithm greedily reassigns each word type to the part-of-speech that results in the greatest increase in likelihood.

In order to incorporate morphology, Clark (2003) associates with each part-of-speech a HMM with letter emissions. The vocabulary is generated by generating a series of word types from the letter HMM of each part-of-speech. These can model very basic concatenative morphology. The parameters of the HMMs are estimated by running a single iteration of Forward-Backward after each round of reassigning words to tags. In our experiments we evaluate both the model without morphology (NE in our experiments), and the morphological model, trying both 5 and 10 states in the letter HMM (NEMorph5, NEMorph10 respectively).

2.3.2 Two-Level HMM

The final part-of-speech inducer we try from Clark (2003) is a two-level HMM. This is similar to the previous model, except it lifts the restriction that a word appear under only one part-of-speech. Alternatively, one could think of this model as a standard HMM, whose emission distributions incorporate a mixture of a letter HMM and a standard multinomial. Training uses a simple variation of Forward-Backward. In the experiments in this paper, we initialize the mixture parameters to .5, and try 5 states in the letter HMM. We refer to this model as 2HMM.

2.4 Tag Evaluation

Objective evaluation in any clustering task is always difficult, since there are many ways to define good clusters. Typically it involves a mixture of subjective evaluation and a comparison of the clusters to those found by human annotators. In the realm of part-of-speech induction, there are several common ways of doing the latter. These split into two groups: accuracy and information-theoretic criteria.

Accuracy, given some mapping between the set of induced classes and the gold standard labels, is the number of words in the corpus that have been marked with the correct gold label divided by the total number of word tokens. The main challenge facing these metrics is deciding how to to map each induced part-of-speech class to a gold tag. One option is what Johnson (2007) calls "many-to-one" (M-to-1) accuracy, in which each induced tag is labeled with its most frequent gold tag. Although this results in a situation where multiple induced tags may share a single gold tag, it does not punish a system for providing tags of a finer granularity than the gold standard.

In contrast, "one-to-one" (1-to-1) accuracy restricts each gold tag to having a single induced tag. The mapping typically is made to try to give the most favorable mapping in terms of accuracy, typically using a greedy assignment (Haghighi and Klein, 2006). In cases where the number of gold tags is different than the number of induced tags, some must necessarily remain unassigned (Johnson, 2007).

In addition to accuracy, there are several information theoretic criteria presented in the literature. These escape the problem of trying to find an appropriate mapping between induced and gold tags, at the expense of perhaps being less intuitive.

Let T_I be the tag assignments to the words in the corpus created by an unsupervised tagger, and let T_G be the gold standard tag assignments. Clark (2003) uses Shannon's conditional entropy of the gold tagging given the induced tagging $H(T_G|T_I)$. Lower entropy indicates less uncertainty in the gold tagging if we already know the induced tagging. Freitag (2004) uses the similar "cluster-conditional tag perplexity" which is merely $exp(H(T_G|T_I))$[2]. Since cluster-conditional tag perplexity is a monotonic function of $H(T_G|T_I)$, we only report the latter.

Goldwater and Griffiths (2007) propose using the Variation of Information of Meilă (2003):

$$VI(T_G; T_I) = H(T_G|T_I) + H(T_I|T_G)$$

VI represents the change in information when going from one clustering to another. It holds the nice properties of being nonnegative, symmetric, as well as fulfilling the triangle inequality.

[1] http://www.cs.rhul.ac.uk/home/alexc/pos.tar.gz

[2] Freitag (2004) measures entropy in nats, while we use bits. The difference is a constant factor.

3 Grammar Induction

In addition to parts-of-speech, we also want to discover deeper syntactic relationships. Grammar induction is the problem of determining these relationships in an unsupervised fashion. This can be thought of more concretely as an unsupervised parsing task. As there are many languages and domains with few treebank resources, systems that can learn syntactic structure from unlabeled data would be valuable. Most work on this problem has focused on either dependency induction, which we discuss in Section 3.2, or on constituent induction, which we examine in the next section.

The Grammar Induction systems we use to evaluate the above taggers are the Constituent-Context Model (CCM), the Dependency Model with Valence (DMV), and a model which combines the two (CCM+DMV) outlined in (Klein and Manning, 2002; Klein and Manning, 2004).

3.1 Constituent Grammar Induction

Klein and Manning (2002) present a generative model for inducing constituent boundaries from part-of-speech tagged text. The model first generates a bracketing $B = \{B_{ij}\}_{1 \leq i \leq j \leq n}$, which specifies whether each span (i, j) in the sentence is a constituent or a distituent. Next, given the constituency or distituency of the span B_{ij}, the model generates the part-of-speech yield of the span $t_i...t_j$, and the one-tag context window of the span t_{i-1}, t_{j+1}. $P(t_i...t_j|B_{ij})$ and $P(t_{i-1}, t_{j+1}|B_{ij})$ are multinomial distributions. The model is trained using EM.

We evaluate induced constituency trees against those of the Penn Treebank using the versions of unlabeled precision, recall, and F-score used by Klein and Manning (2002). These ignore trivial brackets and multiple constituents spanning the same bracket. They evaluate their CCM system on the Penn Treebank WSJ sentences of length 10 or less, using part-of-speech tags induced by the baseline system of Schütze (1995). They report that switching to induced tags decreases the overall bracketing F-score from 71.1 to 63.2, although the recall of VP and S constituents actually improves. Additionally, they find that NP and PP recall decreases substantially with induced tags. They attribute this to the fact that nouns end up in many induced tags.

There has been quite a bit of other work on constituency induction. Smith and Eisner (2004) present an alternative estimation technique for CCM which uses annealing to try to escape local maxima. Bod (2006) describes an unsupervised system within the Data-Oriented-Parsing framework. Several approaches try to learn structure directly from raw text. Seginer (2007) has an incremental parsing approach using a novel representation called common-cover-links, which can be converted to constituent brackets. Van Zaanen (2001)'s ABL attempts to align sentences to determine what sequences of words are substitutable.

The work closest in spirit to this paper is Cramer (2007), who evaluates several grammar induction systems on the Eindhoven corpus (Dutch). One of his experiments compares the grammar induction performance of these systems starting with tags induced using the system described by Biemann (2006), to the performance of the systems on manually-marked tags. However he does not evaluate to what degree better tagging performance leads to improvement in these systems.

3.2 Dependency Grammar Induction

A dependency tree is a directed graph whose nodes are words in the sentence. A directed edge exists between two words if the target word (argument) is a dependent of the source word (head). Each word token may be the argument of only one head, but a head may have several arguments. One word is the head of the sentence, and is often thought of as the argument of a virtual "Root" node.

Klein and Manning (2004) present their Dependency Model with Valence (DMV) for the unsupervised induction of dependencies. Like the constituency model, DMV works from parts-of-speech. Under this model, for a given head, h, they first generate the parts-of-speech of the arguments to the right of h, and then those to the left. Generating the arguments in a particular direction breaks down into two parts: deciding whether to stop generating in this direction, and if not, what part-of-speech to generate as the argument. The argument decision conditions on h and the direction. The stopping decision conditions on this and also on whether h has already generated an argument in this direction, thereby capturing the limited notion of valence from which the model takes its name. It is worth noting that this model can only represent projective dependency trees, i.e. those without crossing edges.

Dependencies are typically evaluated using di-

Tagger	No. Tags	Tagging Metrics				Grammar Induction Metrics					
						CCM	CCM+DMV			DMV	
		1-to-1	$H(T_G\|T_I)$	M-to-1	VI	UF1	DA	UA	UF1	DA	UA
Gold		1.00	0.00	1.00	0.00	71.50	52.90	67.60	56.50	45.40	63.80
HMM-EM	10	0.39	2.67	0.41	4.39	**58.89**	40.12	**59.26**	**59.43**	**36.77**	**57.37**
HMM-EM	20	**0.43**	2.28	0.48	4.54	**57.31**	**51.16**	**64.66**	**61.33**	**38.65**	**58.57**
HMM-EM	50	0.36	1.83	0.58	4.92	**56.56**	**48.03**	**63.84**	**58.02**	**39.30**	**58.84**
HMM-VB	10	0.40	2.75	0.41	4.42	39.05	27.72	52.84	**58.64**	23.94	51.64
HMM-VB	20	0.40	2.63	0.43	4.65	37.60	33.77	55.97	40.30	30.36	51.53
HMM-VB	50	0.38	2.70	0.42	5.01	34.68	37.29	57.72	39.82	29.03	50.50
NE	10	0.34	2.74	0.40	4.32	28.80	20.70	50.60	32.70	26.20	48.90
NE	20	**0.48**	2.02	0.55	**3.76**	32.50	36.00	**59.30**	40.60	32.80	54.00
NEMorph10	10	**0.44**	2.46	0.47	**3.74**	29.03	25.99	53.80	34.58	26.98	48.72
NEMorph10	20	**0.48**	1.94	0.56	**3.65**	31.95	35.85	57.93	38.22	30.45	50.72
NEMorph10	50	**0.47**	1.24	**0.72**	**3.60**	31.07	36.29	57.76	39.28	31.50	52.83
NEMorph5	10	**0.45**	2.50	0.47	**3.76**	29.04	22.72	51.58	32.67	23.62	47.89
NEMorph5	20	**0.44**	2.02	0.56	**3.80**	31.94	24.17	52.43	32.90	22.41	47.17
NEMorph5	50	**0.47**	**1.27**	**0.72**	**3.64**	31.39	38.63	**59.44**	40.23	34.26	**54.63**
2HMM	10	0.38	2.78	0.41	4.55	31.63	36.35	**58.87**	44.97	28.43	49.32
2HMM	20	0.41	2.35	0.48	4.71	42.39	43.91	**60.74**	50.85	29.32	50.69
2HMM	50	0.37	1.92	0.58	5.11	41.18	**49.94**	**64.87**	**57.84**	**39.24**	**59.14**
SVD	10	0.31	3.07	0.34	4.99	37.77	27.64	49.56	36.46	20.74	45.52
SVD	20	0.33	2.73	0.40	4.99	37.17	30.14	51.66	37.66	22.24	46.25
SVD	50	0.34	2.37	0.47	5.18	36.87	37.66	56.49	52.83	22.50	46.52
SVD	100	0.34	2.03	0.53	5.37	45.46	41.68	**58.83**	**64.20**	20.81	44.36
SVD	200	0.32	1.72	0.59	5.59	**61.90**	34.79	52.25	**59.93**	22.66	42.30

Table 1: The performance of the taggers regarding both tag and grammar induction metrics on WSJ sections 0-10, averaged over 10 runs. Bold indicates the result was within 10 percent of the best-scoring induced system for a given metric.

rected and undirected accuracy. These are the total number of proposed edges that appear in the gold tree divided by the total number of edges (the number of words in the sentence). Directed accuracy gives credit to a proposed edge if it is in the gold tree and is in the correct direction, while undirected accuracy ignores the direction.

Klein and Manning (2004) also present a model which combines CCM and DMV into a single model, which we show as CCM+DMV. In their experiments, this model performed better on both the constituency and dependency induction tasks. As with CCM, Klein and Manning (2004) similarly evaluate the combined CCM+DMV system using tags induced with the same method. Again they find that overall bracketing F-score decreases from 77.6 to 72.9 and directed dependency accuracy measures decreases from 47.5 to 42.3 when switching to induced tags from gold. However for each metric, the systems still do quite well with induced tags.

As in the constituency case, Smith (2006) presents several alternative estimation procedures for DMV, which try to minimize the local maximum problems inherent in EM. It is thus possible these methods might yield better performance for the models when run off of induced tags.

4 Experiments

We induce tags with each system on the Penn Treebank Wall Street Journal (Marcus et al., 1994), sections 0-10, which contain 20,260 sentences. We vary the number of tags (10, 20, 50) and run each system 10 times for a given setting. The result of each run is used as the input to the CCM, DMV, and CCM+DMV systems. While the tags are induced from all sentences in the section, following the practice in (Klein and Manning, 2002; Klein and Manning, 2004), we remove punctuation, and consider only sentences of length not greater than 10 in our grammar induction experiments. Taggings are evaluated after punctuation is removed, but before filtering for length.

To explore the relationship between tagging metrics and the resulting performance of grammar induction systems, we examine each pair of tagging and grammar induction metrics. Consider the following two examples: DMV directed accuracy vs. $H(T_G|T_I)$ (Figure 1), and CCM f-score vs. variation of information (Figure 2). These were selected because they have relatively high magnitude τs. From these plots it is clear that although there

may be a slight correspondence, the relationships are weak at best.

Each tagging and grammar induction metric gives us a ranking over the set of taggings of the data generated over the course of our experiments. These are ordered from best to worst according to the metric, so for instance $H(T_G|T_I)$ would give highest rank to its lowest value. We can compare the two rankings using Kendall's τ (see Lapata (2006) for an overview), a nonparametric measure of correspondence for rankings. τ measures the difference between the number of concordant pairs (items the two rankings place in the same order) and discordant pairs (those the rankings place in opposite order), divided by the total number of pairs. A value of 1 indicates the rankings have perfect correspondence, -1 indicates they are in the opposite order, and 0 indicates they are independent. The τ values are shown in Table 2. The scatter-plot in Figure 1 shows the τ with the greatest magnitude. However, we can see that even these rankings have barely any relationship.

An objection one might raise is that the lack of correspondence reflects poorly not on these metrics, but upon the grammar induction systems we use to evaluate them. There might be something about these models in particular which yields these low correlations. For instance these grammar inducers all estimate their models using EM, which can get caught easily in a local maximum.

To this possibility, we respond by pointing to performance on gold tags, which is consistently high for all grammar induction metrics. There is clearly some property of the gold tags which is exploited by the grammar induction systems even in the absence of better estimation procedures. This property is not reflected in the tagging metrics.

The scores for each system for tagging and grammar induction, averaged over the 10 runs, are shown in Table 1. Additionally, we included runs of the SVD-tagger for 100 and 200 tags, since running this system is still practical with these numbers of tags. The Ney-Essen with Morphology taggers perform at or near the top on the various tagging metrics, but not well on the grammar induction tasks on average. HMM-EM seems to perform on average quite well on all the grammar induction tasks, while the SVD-based systems yield the top bracketing F-scores, making use of larger numbers of tags.

Tagging Metrics	Grammar Induction Metrics					
	CCM	CCM+DMV			DMV	
	UF1	DA	UA	UF1	DA	UA
1-to-1	-0.22	-0.04	0.05	-0.13	0.13	0.12
M-to-1	-0.09	0.17	0.24	0.03	0.26	0.25
$H(T_G\|T_I)$	0.01	0.21	0.27	0.07	0.29	0.28
VI	-0.25	-0.17	-0.06	-0.20	0.07	0.07

Table 2: Kendall's τ, between tag and grammar induction criteria.

4.1 Supervised Experiments

One question we might ask is whether these tagging metrics capture information relevant to any parsing task. We explored this by experimenting with a supervised parser, training off trees where the gold parts-of-speech have been removed and replaced with induced tags. Our expectation was that the brackets, the head propagation paths, and the phrasal categories in the training trees would be sufficient to overcome any loss in information that the gold tags might provide. Additionally it was possible the induced tags would ignore rare parts-of-speech such as FW, and make better use of the available tags, perhaps using new distributional clues not in the original tags.

To this end we modified the Charniak Parser (Charniak, 2000) to train off induced parts-of-speech. The Charniak parser is a lexicalized PCFG parser for which the part-of-speech of a head word is a key aspect of its model. During training, the head-paths from the gold part-of-speech tags are retained, but we replace the tags themselves.

We ran experiments using the bitag HMM from Section 2.2 trained using EM, as well as with the Schütze SVD tagger from Section 2.1. The parser was trained on sections 2-21 of the Penn Treebank for training and section 24 was used for evaluation.

As before we calculated τ scores between each tagging metric and supervised f-score. Unlike the unsupervised evaluation where we used the metric UF1, we use the standard EVALB calculation of unlabeled f-score. The results are shown in Table 3.

The contrast with the unsupervised case is vast, with very high τs for both accuracy metrics. Consider f-score vs. many-to-one, plotted in Figure 3. The correspondence here is very clear: taggings with high accuracy do actually reflect on better parser performance. Note, however, that the correspondence between the information theoretic measures and parsing performance is still rather weak.

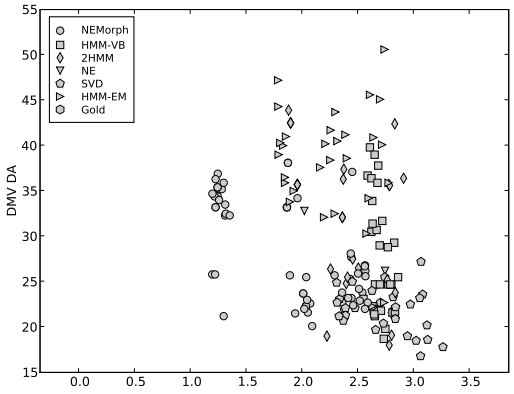

Figure 1: DMV Directed Accuracy vs. $H(T_G|T_I)$

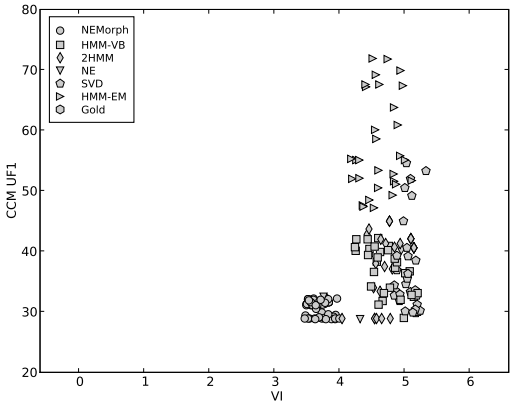

Figure 2: CCM fscore vs. tagging variation of information.

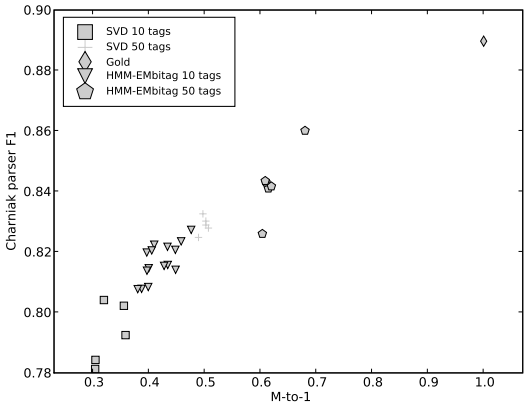

Figure 3: Supervised parsing f-score vs. tagging many-to-one accuracy.

Tagging Metric	Supervised F1
1-to-1	0.62
M-to-1	0.83
$H(T_G\|T_I)$	-0.19
VI	0.25

Table 3: Kendall's τ, between tag induction criteria and supervised parsing unlabeled bracketing F-score.

Interestingly, parsing performance and speed does degrade considerably when training off induced tags. We are not sure what causes this. One possibility is in the lexicalized stage of the parser, where the probability of a head word is smoothed primarily by its part-of-speech tag. This requires that the tag be a good proxy for the syntactic role of the head. In any case this warrants further investigation.

5 Conclusion and Future Work

In this work, we found that none of the most common part-of-speech tagging metrics bear a strong relationship to good grammar induction performance. Although our experiments only involve English, the poor correspondence we find between the various tagging metrics and grammar induction performance raises concerns about their relationship more broadly. We additionally found that while tagging accuracy measures do correlate with better supervised parsing, common information theoretic ones do not strongly predict better performance on either task. Furthermore, the supervised experiments indicate that informative part-of-speech tags are important for good parsing.

The next step is to explore better tagging metrics that correspond more strongly to better grammar induction performance. A good metric should use all the information we have, including the gold trees, to evaluate. Finally, we should explore grammar induction schemes that do not rely on prior parts-of-speech, instead learning them from raw text at the same time as deeper structure.

Acknowledgments

We thank Dan Klein for his grammar induction code, as well as Matt Lease and other members of BLLIP for their feedback. This work was partially supported by DARPA GALE contract HR0011-06-2-0001 and NSF award 0631667.

References

Baum, L.E. 1972. An inequality and associated maximization techniques in statistical estimation of probabilistic functions of Markov processes. *Inequalities*, 3:1–8.

Biemann, Chris. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*, pages 7–12, Sydney, Australia.

Bod, Rens. 2006. An all-subtrees approach to unsupervised parsing. In *Proceedings of Coling/ACL 2006*, pages 865–872.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the North American Chapter of the ACL 2000*, pages 132–139.

Clark, Alexander. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of EACL 2003*, pages 59–66, Budapest, Hungary.

Cramer, Bart. 2007. Limitations of current grammar induction algorithms. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 43–48.

Dasgupta, Sajib and Vincent Ng. 2007. Unsupervised part-of-speech acquisition for resource-scarce languages. In *Proceedings of the EMNLP/CoNLL 2007*, pages 218–227.

Freitag, Dayne. 2004. Toward unsupervised whole-corpus tagging. In *Proceedings of Coling 2004*, pages 357–363, Aug 23–Aug 27.

Goldwater, Sharon and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the ACL 2007*, pages 744–751.

Haghighi, Aria and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of HLT/NAACL 2006*, pages 320–327, New York, USA.

Johnson, Mark. 2007. Why doesn't EM find good HMM POS-taggers? In *Proceedings of the EMNLP/CoNLL 2007*, pages 296–305.

Klein, Dan and Christopher Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of ACL 2002*.

Klein, Dan and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of ACL 2004*, pages 478–485, Barcelona, Spain, July.

Lapata, Mirella. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):1–14.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*.

Meilă, Marina. 2003. Comparing clusterings. *Proceedings of the Conference on Computational Learning Theory (COLT)*.

Merialdo, Bernard. 1994. Tagging english text with a probabilistic model. *Computational Linguistics*, 20(2):154–172.

Ney, Herman, Ute Essen, and Renhard Knesser. 1994. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38.

Schütze, Hinrich. 1995. Distributional part-of-speech tagging. In *Proceedings of the 7th conference of the EACL*, pages 141–148.

Seginer, Yoav. 2007. Fast unsupervised incremental parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 384–391, Prague, Czech Republic.

Smith, Noah A. and Jason Eisner. 2004. Annealing techniques for unsupervised statistical language learning. In *Proceedings of ACL 2004*, pages 487–494, Barcelona, Spain.

Smith, Noah A. and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of ACL 2005*, pages 354–362, Ann Arbor, Michigan.

Smith, Noah A. 2006. *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. Ph.D. thesis, Department of Computer Science, Johns Hopkins University, October.

Van Zaanen, Menno M. 2001. *Bootstrapping Structure into Language: Alignment-Based Learning*. Ph.D. thesis, University of Leeds, September.