
Unsupervised Part of Speech Tagging Without a Lexicon

Adam R. Teichert
School of Computing
University of Utah
Salt Lake City, UT 84112
teichert@cs.utah.edu

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, UT 84112
hal@cs.utah.edu

1 Introduction

As digital data and computing power become less expensive and as the premium on human time climbs ever higher, research in natural language processing has turned toward unsupervised methods of solving some of its most fundamental tasks including part of speech (POS) tagging and dependency parsing.

Unsupervised dependency parsing frequently assume that input sentences have already been labeled with POS tags. Likewise, most unsupervised POS taggers (including those proposed by [1] and [2]) either produce numeric labels on words without providing a mapping to POS tags or they rely on language specific lexical information such as lists reporting the possible tags that some or all of the words can take. However, linguists have devoted decades of research toward identifying features of word order in various languages and toward understanding principles that influence the structure of natural languages in general [3] [4].

We suggest two lexicon independent sources for prior information in unsupervised POS tagging. First we discuss the notion of class “openness” cues and show dramatic improvements in unsupervised POS tagging performance over a basic HMM style model on a Portuguese dataset. Secondly, we review the linguistic notion of language word order features and show results that suggest that using such rules can allow an unsupervised POS tagger to correctly assign actual POS labels (rather than arbitrary numeric labels) to at least some word classes without any lexicon or prior knowledge of any words in the vocabulary.

2 Related Work

Before presenting our models, we list a few interesting deviations from the typical dependence on POS tags for unsupervised dependency parsing and the significant lexical information for POS tagging that we refer to in the introduction. [5] attain good results by parsing directly at the text level rather than the POS level and so avoid the need for POS tags in parsing. Others have reported results of unsupervised parsers on automatically induced syntactic clusters of words. In unsupervised POS tagging, [6], [7], and others show the effects on accuracy of varying the amount of dictionary information given to the tagger. [8] presents an approach that reduces his tagger’s dependence on lexical information. None of these, however, attempt to disambiguate POS classes with no lexical prior information.

3 Models and Motivation

We begin with a generative story for a basic HMM approach to unsupervised POS tagging. The first tag of the sentence is generated given that it is the first tag of the sentence, and the first word is emitted given the generated tag. Each subsequent tag is then generated in turn given the tag

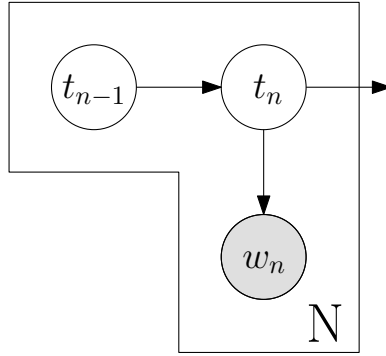


Figure 1: HMM graphical model.

immediately preceding it, and for each generated tag, a word is emitted given that tag. Figure 1 shows the basic HMM graphical model.

POS tagging can be seen as a clustering of word instances from a document into groups that share a particular POS tag. Inferring POS tags using the HMM structure described above encourages each cluster to primarily contain instances of just a few word types while encouraging only a few tags to have high probability of following any given tag.

We now discuss our contribution of two additional categories of linguistic properties that lead to improvements to this basic model.

3.1 Tag Openness

Some parts of speech, like nouns, verbs, adjectives, and adverbs are constantly being expanded as words are created in the language. Other parts of speech, like articles, pronouns, and prepositions are basically closed, and it would be rare for people to stumble upon a new closed-class word in their native language that they have not seen before.

Motivated by this observation, we propose the following three additional properties that we would like to find in a POS clustering of word instances.

1. Some clusters should represent closed-class tags and some should represent open-class tags
2. In clusters representing closed-class tags we expect to find only a few word types and a disproportionately large number of word instances
3. In clusters representing open-class tags we expect to find a large variety of word types and a more proportionate number of word instances

To encourage these properties, we propose the modified graphical model shown in Figure 2. The basic HMM already imposes the restriction that we decide a priori on the number (T) of clusters (POS tags) with which we will label the data. Our model now includes T additional nodes which can take binary values representing the openness of the tag to which it corresponds. Our generative story is now modified such that, a word is generated given its tag and the openness of the tag. In addition to affecting which tag distribution the word is sampled from, the smoothing parameter on closed-class words is smaller than for open-class words since we do not expect to see “unfamiliar” closed-class words. In our experiments, we treat the openness of clusters as observed and we allocate half of the clusters as closed-class and half as open-class.

Finally, we postulate that the openness of a given node will have predictive value in guessing what the openness of the next node will be. To model this, we introduce an additional openness node for each word node. The value of this openness node is generated given the openness of the previous node, and we employ a product-of-experts to generate a word given its tag, its openness, and the assumed openness of its tag. Specifically if t_n is the tag on the n^{th} word, o_n is the local predication of its openness, and g_1, \dots, g_T are the global assumptions of the openness of the corresponding tags, then $P(w_n|t_n, o_n, g_1, \dots, g_T) = P(w_n|t_n, o_n, g_{t_n})$ is factored as $P(w_n|t_n, o_n)P(w_n|t_n, g_{t_n})$.

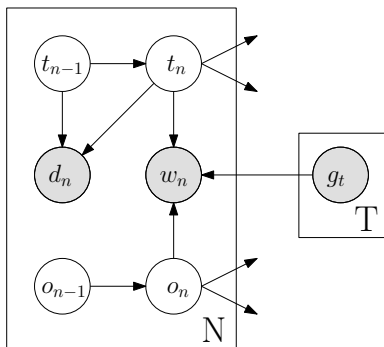


Figure 2: HMM graphical model with openness.

3.2 Language Word Order Features and Cluster Labeling

While we expect the principles of tag openness to be generally applicable across languages, we now explore another category of linguistically motivated prior information which is language dependent, but which may be easily available for many languages and perhaps inferred for others. This category of information includes language word order tendency features. The following two examples of this type of feature apply both to English and to Portuguese:

1. Since English has prepositions rather than postpositions, in an adpositional phrase, such as “to the old store”, the adposition (“to”) tends to precede the noun of the phrase (“store”)
2. If an article is used in referring to a noun (for example “the angry dog”), the article (“the”) tends to come before the noun that it modifies (“dog”)

Although we will see that modeling tag openness helps to dramatically improve the model’s ability to cluster the word instances into POS tag groups, we are interested in a way for us to identify which of the resulting clusters corresponds to nouns, to articles, or to prepositions. In general, however, previous work has not attempt to perform this identification without using lexical information such as labeled seed words or partial tag dictionaries. Using the ordering information encoded in these rules we extend our model in an attempt to identify the three parts of speech used in our two rules above (i.e. *noun*, *article*, and *preposition*).

An alternative to the basic HMM generative story is to begin with building a tree that reflects how many words will be generated in the sentence and how those words will relate syntactically to one another (i.e.. the structure of a dependency parse). Tags and words are then generated as in the basic HMM case except that we start with the root node of the tree and children tags are generated given parent tags; as in the basic HMM case, each word is then generated given its tag.

Although it may seem odd to assume a correct dependency parse tree when undertaking unsupervised POS tagging, we note that unsupervised parsers often assume correct POS tags to infer a tree structure and that our approach merely works on the other half of the same problem, inferring the POS tags given the tree structure. We leave to further work experiments in bootstrapping this process or in the joint inference of both tags and trees, but we note that unsupervised dependency parsers exist which do not rely on named POS labels [9], so the following is one potential bootstrapping approach :

1. Use HMM based unsupervised POS tagger to assign numbers as POS tags to word instances (without using the dependency tree structure)
2. Use unsupervised dependency parser on current tagging to infer a dependency parse structure over the sentences
3. Use tree based unsupervised POS tagger to assign POS tags to word instances
4. If not converged, repeat from (2)

Assuming a given tree structure, we can now know, for example, that the parent of the word “the” in the example above is the word “dog”. Furthermore, we can see that having the child on the left

of the parent is consistent with a “noun” label for “dog” and an “article” label for “the”. Having the child on the right of the parent would not be consistent with that tagging.

We now modify the tree HMM generative story so that the tree is built as tags are generated. In particular, as each child node in the tree is created and its tag is generated, the decision of whether to place the child to the left or to the right of the parent node is made according to some probability distribution, $P(\text{Direction}|\text{Tag}, \text{ParentTag})$. Since the direction variable in our new model is observed, the true direction will be able to work with our two word order rules to influence the tagging and to identify the three POS tags used in our rules. Figure 3 shows the final graphical model.

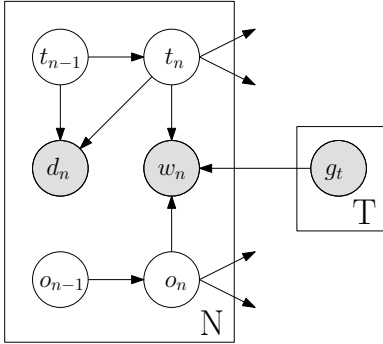


Figure 3: Tree HMM graphical model with openness and direction.

We now encode our two rules by arbitrarily picking a distinct tag id for each of the three parts of speech in our rules and assigning a steep prior probability for $P(\text{Direction} = \text{left}|\text{Tag} = \text{article}, \text{ParentTag} = \text{noun})$ and $P(\text{Direction} = \text{right}|\text{Tag} = \text{noun}, \text{ParentTag} = \text{preposition})$.

The final piece of information that we give our model is that the tag id named “noun” should have a global openness value of “open”, and that the tag ids we named “preposition” and “article” should each have a global openness value of “closed”.

4 Experiments and Results

Experiments on multiple languages and tag sets would certainly be appropriate in evaluating our model changes and the usefulness our word-order rules and the openness information. In this abstract, however, we focus only on the their effect on the Portuguese portion of the CONLL 2007 [10], [11] training data which consists of 206678 tokens divided among 16 tags (however the four least frequent tags occur only 48 times in total). Nouns are the most frequent tag representing $\approx 18.84\%$ of the dataset.

To visualize the effects of our model changes on classification accuracy, Figure 4 shows the tagging accuracy over the first 200 iterations of Gibbs sampling using the various models. The accuracies reported are found by using the gold tagging to align each of our clusters with exactly one of the POS tags in the gold tag set. Note, the alignment is optimal in the sense that there is no single swap of cluster labels that would result in an alignment yielding a higher accuracy. Each curve is the average accuracy over three runs.

In Figure 4, we see that adding the openness information to the model helps the basic model significantly. Also, while the direction rules do not appear to be helping with clustering accuracy, Figure 5 shows the main contribution of these rules, comparing the two models which incorporate rules with the baseline of labeling everything as a noun. As discussed, one of the unique features of this work is our ability to identify some of the clusters as representing particular POS tags and the direction rules give the model information that can help in this identification. For the curves in Figure 5, word instances which were labeled with POS tags that did not have relevant direction rules, were counted as mistagged. The curve labeled “extra rules” incorporates direction rules in addition to the two that we discussed earlier.

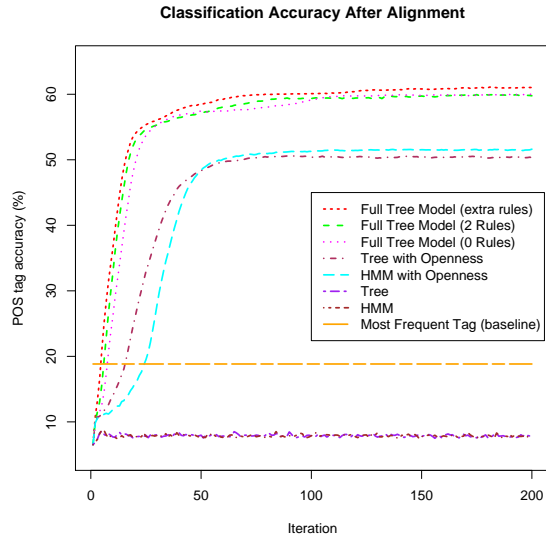


Figure 4: Comparison of models after swapping tags to find best correspondence.

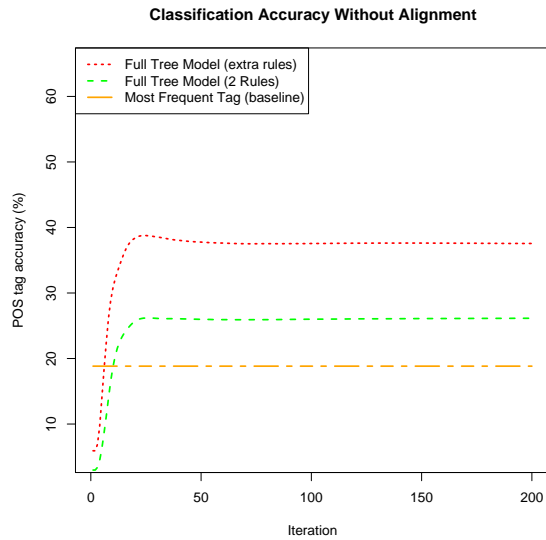


Figure 5: Comparison of direction rule models without swapping tags.

5 Conclusion

While our results are not meant to compete with the state of the art, we have shown that incorporating the notion of POS tag openness into an unsupervised POS tagger dramatically improves its ability to cluster word instances in our Portuguese dataset into POS tag clusters. We also have proposed the use of language word-order features in conjunction with the structure of the dependency parse tree as a source of further prior knowledge and as a novel method of assigning POS names to unsupervised POS clusters without relying on lexicon specific information.

References

- [1] A. Clark. Inducing syntactic categories by context distribution clustering. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 91–94. Association for Computational Linguistics Morristown, NJ, USA, 2000.
- [2] E. Brill. Unsupervised learning of disambiguation rules for part of speech tagging. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 1–13, 1995.
- [3] J. Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. 1963, pages 73–113, 1963.
- [4] M. Dryer. The greenbergian word order correlations. *Language*, 68(1):81–138, 1992.
- [5] Y. Seginer. Fast unsupervised incremental parsing. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 384, 2007.
- [6] S. Goldwater and T. Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, page 744751, 2007.
- [7] B. Snyder, T. Naseem, J. Eisenstein, and R. Barzilay. Adding more languages improves unsupervised multilingual part-of-speech tagging: A Bayesian non-parametric approach. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on ZZZ*, pages 83–91. Association for Computational Linguistics, 2009.
- [8] Q. Zhao and M. Marcus. A Simple Unsupervised Learner for POS Disambiguation Rules Given Only a Minimal Lexicon. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.
- [9] S. Cohen, K. Gimpel, and N. Smith. Logistic normal priors for unsupervised probabilistic grammar induction. In *Proceedings of NAACL-HLT*, 2009.
- [10] J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, volume 7, pages 915–932, 2007.
- [11] S. Afonso, E. Bick, R. Haber, and D. Santos. “Floresta sintá(c)tica”: a treebank for Portuguese. In *Proceedings of LREC 2002*, pages 1698–1703, 2002.