Proceedings of HLT-NAACL 2003
Short Papers , pp. 79-81
Edmonton, May-June 2003

Bayesian Nets in Syntactic Categorization of Novel Words

Leonid Peshkin
Dept. of Computer Science
Harvard University
Cambridge, MA
pesha@eecs.harvard.edu

Avi Pfeffer
Dept. of Computer Science
Harvard University
Cambridge, MA
avi@eecs.harvard.edu

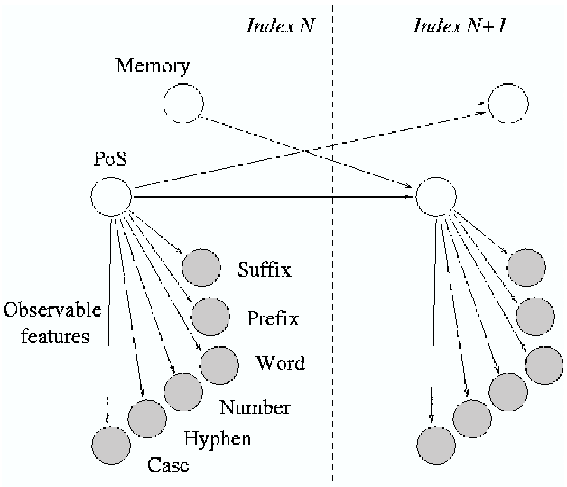
Virginia Savova
Dept. of Cognitive Science
Johns Hopkins University
Cambridge, MA
savova@jhu.edu

Abstract

This paper presents an application of a Dynamic Bayesian Network (DBN) to the task of assigning Part-of-Speech (PoS) tags to novel text. This task is particularly challenging for non-standard corpora, such as Internet lingo, where a large proportion of words are unknown. Previous work reveals that PoS tags depend on a variety of morphological and contextual features. Representing these dependencies in a DBN results into an elegant and effective PoS tagger.

1 Introduction

Uncovering the syntactic structure of texts is a necessary step towards extracting their meaning. In order to obtain an accurate parse for an unseen text, we need to assign Part-of-Speech (PoS) tags to a string of words. This paper covers one aspect of our work of PoS tagging with Dynamic Bayesian Networks (DBNs), which demonstrates their success at tagging unknown (OoV) words. Please refer to the companion paper [Peshkin, 2003] for substantial discussion of our method and other details. Although currently existing algorithms exhibit high word-level accuracy, PoS tagging is not a solved problem. First, even a small percentage of errors may derail subsequent processing steps. Second, the results of tagging are not robust if a large proportion of words are unknown, or if the testing corpus differs in style from the training corpus. At the same time, diverse training corpora are lacking and most taggers are trained on a large annotated corpus extracted from the Wall Street Journal (WSJ). These factors significantly hamper the use PoS tagging to extract information from non-standard corpora, such as email messages and websites. Our work on Information Extraction from an email corpus left us searching for a PoS tagger that would perform well on Internet texts and integrate easily into a large probabilistic reasoning system by producing a distribution over tags rather than deterministic answer. Internet sources exhibit a set of idiosyncratic characteristics not present in the training corpora available to taggers to date. They are often written in telegraphic style, omitting closed-class words, which leads to a higher percentage of ambiguous items. Most importantly, as a consequence of the rapidly evolving Netlingo, Internet texts are full of new words, misspelled words and one-time expressions. These characteristics are bound to lower the accuracy of existing taggers. A look at the literature confirms that error rates for unknown words are quite high. According to several recent publications [Toutanova 2002, Lafferty et al.2002] OoV tagging presents a serious challenge to the field. The transformation-based Brill tagger, achieves 96.5% accuracy for the WSJ, but a mere 85% on unknown words. Existing probabilistic taggers also don't fare well on unknown words. Reported results on OoV rarely exceed Brill's performance by a tiny fraction. They are mostly based on (Hidden) Markov Models [Brants 2000, Kupiec, 1992]. A model based on Conditional Random Fields [Lafferty et al.] outperforms the HMM tagger on unknown words yielding 24% error rate. The best result known to us is achieved by Toutanova[2002] by enriching the feature representation of the MaxEnt approach [Ratnaparkhi, 1996].



2 A DBN for PoS Tagging

Unlike Toutanova[2002], we deliberately base our model on the original feature set of Ratnaparkhi's MaxEnt. Our Bayesian network includes a set of binary features (1-3, below) and a set of vocabulary features (4-6, below). The binary features indicate the presence or absence of a particular character in the token: 1. does the token contain a **capital** letter; 2. does the token contain a **hyphen;** 3. does the token contain a **number**. We used Ratnaparkhi's vocabulary lists to encode the values of 6458 frequent **Word**s, 3602 **Prefix**es and 2925 **Suffix**es up to 4 letters long.

A Dynamic Bayesian network (DBN) is a Bayesian network unwrapped in time, such that it can represent dependencies between variables at adjacent positions (see figure). For a good overview of DBNs, see Murphy [2002]. The set of observable variables in our network consists of the binary and vocabulary features mentioned above. In addition, there are two hidden variables: PoS and Memory which reflects contextual information about past PoS tags. Unlike Ratnaparkhi we do not directly consider any information about preceding words even the previous one [Toutanova 2002]. However, a special value of Memory indicates whether we are at the beginning of the sentence.

Learning in our model is equivalent to collecting statistics over co-occurrences of feature values and tags. This is implemented in GAWK scripts and takes minutes on the WSJ training corpus. Compare this to laborious Improved Iterative Scaling for MaxEnt. Tagging is carried out by the standard Forward-Backward algorithm (see e.g. Murphy[2002]). We do not need to use specialized search algorithms such as Ratnaparkhi's "beam search". In addition, our method does not require a "Development" stage.

Following established data split we use sections (0-22) of WSJ for training and the rest (23-24) as a test set. The test sections contain 4792 sentences out of about 55600 total sentences in WSJ corpus. The average length of a sentence is 23 tokens. In addition, we created two specialized testing corpora (available upon request for comparison purposes). A small Email corpus was prepared from excerpts from the MUC seminar announcement corpus. "The Jabberwocky" is a poem by Louis Carol where the majority of words are made-up, but their PoS tags are apparent to speakers of English. We use "The Jabberwocky" to illustrate performance on unknown words. Both the Email corpus and the Jabberwocky were pre-tagged by the Brill tagger then manually corrected.

We began our experiments by using the original set of features and vocabulary lists of Ratnaparkhi for the variables Word, Prefix and Suffix. This produced a reasonable performance. While investigating the relative contribution of each feature in this setting, we discovered that the removal of the three binary features from the feature set does not significantly alter performance. Upon close examination, the vocabularies we used turned out to contain a lot of redundant information that is otherwise handled by these features. For example, Prefix list contained 84 hyphens (e.g. both "co-" and "co"), 530 numbers and 150 capitalised words, including capital letters. We proceed, using reduced vocabularies obtained by removing redundant information from the original lists. The results are presented in Table 1 for various testing conditions. Since Toutanova[2002] report that Prefix information worsens performance, we conducted the second set of experiments with a network that contained no information about prefix. We found no significant change in performance.

Our overall performance is comparable to the best result known on this benchmark (e.g. Toutanova[2002]. At the same time, our performance on OoV words is significantly better (9.4% versus 13.3%). We attribute this difference to the purer representation of morphologically relevant suffixes in our factored vocabulary, which excludes redundant and therefore potentially confusing information. Another reason may be that our method puts a greater emphasis on the syntactically relevant

facts, such as morphology and tag sequence information by *refraining* to use word-specific cues. Despite our good performance on the WSJ corpus, we failed to improve Brill's tagging on our two specialized corpora. Both Brill and our method achieved 89% on the Jabberwocky poem. Note, however, that Brill uses much more sophisticated mechanisms to obtain this result. It was particularly disappointing for us to find out that we did not succeed in labeling the Email corpus accurately (16.3% versus 14.9% of Brill). However, the reason for this poor performance appears to be partly related to a labeling convention of the Penn Treebank, which essentially causes most capitalized words to be categorized as NNPs. In our view, there is a significant difference between the grammatical status of a proper name "Virginia Savova", where words can't be said to modify one another, and a name of an institution such as "Department of Chemical Engineering", where "chemical" clearly modifies "engineering". While a rule-based system profits from this simplistic convention, our method is harmed by it.

3 Conclusion

Our approach shows promise as it is both probabilistic and outperforms existing statistical taggers on unknown words. We are especially encouraged by our performance on the WSJ and take this as evidence that our method has the potential to significantly improve PoS tagging of non-standard texts. In addition, our method has the advantage of being conceptually simple, fast, and flexible with respect to feature representation. We are currently investigating the performance of other DBN topologies on PoS tagging.

Description	Average %%	OoV Words	Sentence
Original feature set of Ratnaparkhi	6.8	13.2	69.4
Email corpus	16.3	12.2	79.0
Jabberwocky	11.0	23.0	65.0
Trained on WSJ tested on Brown	13.1	26.5	73.2
Factored feature Set on random WSJ	3.6	9.8	52.7

References

Brants, T. 2000. *TnT -- a statistical part-of-speech tagger.* In Proceedings of the 6th ANLP.

Brill. E. 1995. *Transformation-based error-driven learning and natural language processing*. Computational Linguistics, 21(4):543--565.

Charniak, E., C. Hendrickson, N. Jacobson, and M. Perkowitz. 1993. *Equations for part-of-speech tagging*. In Proceedings of 11th AAAI, 1993

Jelinek. F. 1985. Markov source modeling of text generation. In J. K. Skwirzinski, ed., Impact of Processing Techniques on Communication, Dordecht

Kupiec. J. M. 1992, Robust part-of-speech tagging using a hidden Markov model. Computer Speech and Language, 6:225-242.

Lafferty J., McCallum A., Pereira F., *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, Proc. 18th ICML, 2002

Marcus M., G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In ARPA Human Language Technology Workshop.

Manning C. and H. Schutze. 1999. Foundations of Statistical Natural Language Processing. The MIT Press. Cambridge, Massachusetts.

Murphy. K. *Dynamic Bayesian Networks:* Representation, Inference and Learning. PhD thesis. UC Berkeley. 2002.

Peshkin. L., 2003. *Part-of-Speech Tagging with Dynamical Bayesian Network.* manuscript

Peshkin. L., Pfeffer, A. 2003. *Bayesian Information Extraction Network*, manuscript, 2003

Ratnaparkhi. A *A maximum entropy model for part-of-speech tagging*. In Proceedings of EMNLP, 1996.

Samuelsson. C. Morphological Tagging Based Entirely on Bayesian Inference. In 9th Nordic Conference on Computational Linguistics, Stockholm University, Stockholm, Sweden. 1993.

Toutanova K and Manning, C. *Enriching the Knowledge Sources Used in a Maximum Entropy PoS Tagger.* 2002.