Part of Speech Tagging in Context

Michele BANKO and Robert C. MOORE
Microsoft Research
One Microsoft Way
Redmond, WA 98052 USA
{mbanko, bobmoore}@microsoft.com

Abstract

We present a new HMM tagger that exploits context on both sides of a word to be tagged, and evaluate it in both the unsupervised and supervised case. Along the way, we present the first comprehensive comparison of unsupervised methods for part-of-speech tagging, noting that published results to date have not been comparable across corpora or lexicons. Observing that the quality of the lexicon greatly impacts the accuracy that can be achieved by the algorithms, we present a method of HMM training that improves accuracy when training of lexical probabilities is unstable. Finally, we show how this new tagger achieves state-of-the-art results in a supervised, non-training intensive framework.

1 Introduction

The empiricist revolution in computational linguistics has dramatically shifted the accepted boundary between what kinds of knowledge are best supplied by humans and what kinds are best learned from data, with much of the human-supplied knowledge now being in the form of annotations of data. As we look to the future, we expect that relatively unsupervised methods will grow in applicability, reducing the need for expensive human annotation of data.

With respect to part-of-speech tagging, we believe that the way forward from the relatively small number of languages for which we can currently identify parts of speech in context with reasonable accuracy will make use of unsupervised methods that require only an untagged corpus and a lexicon of words and their possible parts of speech. We believe this based on the fact that such lexicons exist for many more languages (in the form of conventional dictionaries) than extensive human-tagged training corpora exist for.

Unsupervised part-of-speech tagging, as defined above, has been attempted using a variety of learning algorithms (Brill 1995, Church, 1988, Cutting et. al. 1992, Elworthy, 1994 Kupiec 1992, Merialdo 1991). While this makes unsupervised part-of-speech tagging a relatively well-studied problem, published results to date have not been comparable with respect to the training and test data used, or the lexicons which have been made available to the learners.

In this paper, we provide the first comprehensive comparison of methods for unsupervised part-of-speech tagging. In addition, we explore two new ideas for improving tagging accuracy. First, we explore an HMM approach to tagging that uses context on both sides of the word to be tagged, inspired by previous work on building bidirectionality into graphical models (Lafferty et. al. 2001, Toutanova et. al. 2003). Second we describe a method for sequential unsupervised training of tag sequence and lexical probabilities in an HMM, which we observe leads to improved accuracy over simultaneous training with certain types of models.

In section 2, we provide a brief description of the methods we evaluate and review published results. Section 3 describes the contextualized variation on HMM tagging that we have explored. In Section 4 we provide a direct comparison of several unsupervised part-of-speech taggers, which is followed by Section 5, in which we present a new method for training with suboptimal lexicons. In section 6, we revisit our new approach to HMM tagging, this time, in the supervised framework.

2 Previous Work

A common formulation of an unsupervised part-of-speech tagger takes the form of a hidden Markov model (HMM), where the states correspond to part-of-speech tags, t_i, and words, w_i, are emitted each time a state is visited. The training of HMM–based taggers involves estimating lexical probabilities, $P(w_i|t_i)$, and tag sequence probabilities, $P(t_i \mid t_{i-1} \ldots t_{i-n})$. The ultimate goal of HMM training is to find the model that maximizes the probability of a given training text, which can be done easily using the forward-backward, or Baum-Welch algorithm (Baum et al 1970, Bahl, Jelinek and Mercer, 1983). These model probabilities are then used in conjunction with the Viterbi algorithm (Viterbi, 1967) to find the most probable sequence of part-of-speech tags for a given sentence.

When estimating tag sequence probabilities, an HMM tagger, such as that described in Merialdo

(1991), typically takes into account a history consisting of the previous two tags -- e.g. we compute $P(t_i \mid t_{i-1}, t_{i-2})$. Kupiec (1992) describes a modified trigram HMM tagger in which he computes word classes for which lexical probabilities are then estimated, instead of computing probabilities for individual words. Words contained within the same equivalence classes are those which possess the same set of possible parts of speech.

Another highly-accurate method for part-of-speech tagging from unlabelled data is Brill's unsupervised transformation-based learner (UTBL) (Brill, 1995). Derived from his supervised transformation-based tagger (Brill, 1992), UTBL uses information from the distribution of unambiguously tagged data to make informed labeling decisions in ambiguous contexts. In contrast to the HMM taggers previously described, which make use of contextual information coming from the left side only, UTBL considers both left and right contexts.

Reported tagging accuracies for these methods range from 87% to 96%, but are not directly comparable. Kupiec's HMM class-based tagger, when trained on a sample of 440,000 words of the original Brown corpus, obtained a test set accuracy of 95.7%. Brill assessed his UTBL tagger using 350,000 words of the Brown corpus for training, and found that 96% of words in a separate 200,000-word test set could be tagged correctly. Furthermore, he reported test set accuracy of 95.1% for the UTBL tagger trained on 120,000 words of Penn Treebank and tested on a separate test set of 200,000 words taken from the same corpus. Finally, using 1 million words from the Associated Press for training, Merialdo's trigram tagger was reported to have an accuracy of 86.6%. This tagger was assessed using a tag set other than that which is employed by the Penn Treebank.

Unfortunately none of these results can be directly compared to the others, as they have used different, randomized and irreproducible splits of training and test data (Brill and Kupiec), different tag sets (Merialdo) or different corpora altogether.

The HMM taggers we have discussed so far are similar in that they use condition only on left context when estimating probabilities of tag sequences. Recently, Toutanova et al. (2003) presented a supervised conditional Markov Model part-of-speech tagger (CMM) which exploited information coming from both left and right contexts. Accuracy on the Penn Treebank using two tags to the left as features in addition to the current tag was 96.10%. When using tag to the left and tag to the right as features in addition to the current tag, accuracy improved to 96.55%.

Lafferty et al. (2001) also compared the accuracies of several supervised part-of-speech tagging models, while examining the effect of directionality in graphical models. Using a 50%-50% train-test split of the Penn Treebank to assess HMMs, maximum entropy Markov models (MEMMs) and conditional random fields (CRFs), they found that CRFs, which make use of observation features from both the past and future, outperformed HMMs which in turn outperformed MEMMs.

3 Building More Context into HMM Tagging

In a traditional HMM tagger, the probability of transitioning into a state representing tag t_i is computed based on the previous two tags t_{i-1} and t_{i-2}, and the probability of a word w_i is conditioned only on the current tag t_i. This formulation ignores dependencies that may exist between a word and the part-of-speech tags of the words which precede and follow it. For example, verbs which subcategorize strongly for a particular part-of-speech but can also be tagged as nouns or pronouns (e.g. "*thinking* that") may benefit from modeling dependencies on future tags.

To model this relationship, we now estimate the probability of a word w_i based on tags t_{i-1} and t_{i-+1}. This change in structure, which we will call a contextualized HMM, is depicted in Figure 1. This type of structure is analogous to context-dependent phone models used in acoustic modeling for speech recognition (e.g.Young, 1999, Section 4.3).

3.1 Model Definition

In order to build both left and right-context into an HMM part-of-speech tagger, we reformulate the

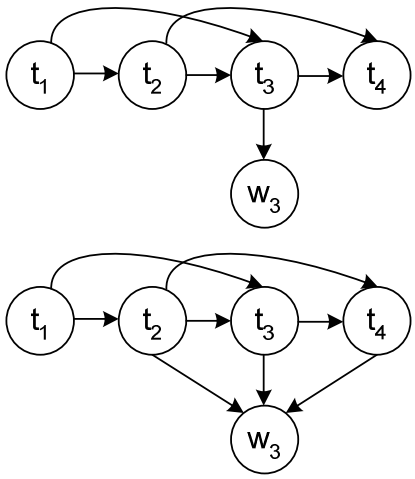

Figure 1: Graphical Structure of Traditional HMM Tagger (top) and Contextualized HMM Tagger (bottom)

trigram HMM model traditionally described as

$$p(W,T) = \prod_{i=1}^{n} p(w_i \mid w_i t_1 .. w_{i-1} t_{i-1} t_i) \times p(t_i \mid w_i t_i .. w_{i-1} t_{i-1})$$

by replacing the approximation:

$$p(w_i \mid w_i t_1 .. w_{i-1} t_{i-1}) = p(w_i \mid t_i)$$
$$p(t_i \mid w_i t_i .. w_{i-1} t_{i-1}) = p(t_i \mid t_{i-2} t_{i-1})$$

with the approximation:

$$p(w_i \mid w_i t_1 .. w_{i-1} t_{i-1}) = p(w_i \mid t_{i-1} t_i t_{i+1})$$
$$p(t_i \mid w_i t_i .. w_{i-1} t_{i-1}) = p(t_i \mid t_{i-2} t_{i-1})$$

Given that we are using an increased context size during the estimation of lexical probabilities, thus fragmenting the data, we have found it desirable to smooth these estimates, for which we use a standard absolute discounting scheme (Ney, Essen and Knesser, 1994).

4 Unsupervised Tagging: A Comparison

4.1 Corpora and Lexicon Construction

For our comparison of unsupervised tagging methods, we implemented the HMM taggers described in Merialdo (1991) and Kupiec (1992), as well as the UTBL tagger described in Brill (1995). We also implemented a version of the contextualized HMM using the type of word classes utilized in the Kupiec model. The algorithms were trained and tested using version 3 of the Penn Treebank, using the training, development, and test split described in Collins (2002) and also employed by Toutanova et al. (2003) in testing their supervised tagging algorithm. Specifically, we allocated sections 00-18 for training, 19-21 for development, and 22-24 for testing. To avoid the problem of unknown words, each learner was provided with a lexicon constructed from tagged versions of the full Treebank. We did not begin with any estimates of the likelihoods of tags for words, but only the knowledge of what tags are possible for each word in the lexicon, i.e., something we could obtain from a manually-constructed dictionary.

4.2 The Effect of Lexicon Construction on Tagging Accuracy

To our surprise, we found initial tag accuracies of all methods using the full lexicon extracted from the Penn Treebank to be significantly lower than previously reported. We discovered this was due to several factors.

One issue we noticed which impacted tagging accuracy was that of a frequently occurring word being mistagged during Treebank construction, as shown in the example in Figure 2a. Since we are not starting out with any known estimates for probabilities of tags given a word, the learner considers this tag to be just as likely as the word's other, more probable, possibilities. In another, more frequently occurring scenario, human annotators have chosen to tag all words in multi-word names, such as titles, with the proper-noun tag, NNP (Figure 2b). This has the effect of adding noise to the set of tags for many closed-class words.

(a) The/VB Lyneses/NNP ,/, of/IN Powder/NNP Springs/NNP ,/, Ga./NNP ,/, have/VBP filed/VBN suit/NN in/IN Georgia/NNP state/NN court/NN against/IN Stuart/NNP James/NNP ,/, *-1/-NONE- alleging/VBG fraud/NN ./.
(b) Last/JJ week/NN CBS/NNP Inc./NNP cancelled/VBD ``/`` The/NNP People/NNP Next/NNP Door/NNP ./. "/"
(c) a/SYM -/: Discounted/VBN rate/NN ./.

Figure 2: Manually-Tagged Examples

Finally, we noticed that a certain number of frequently occurring words (e.g. *a*, *to*, *of*) are sometimes labeled with infrequently occurring tags (e.g. *SYM*, *RB*), as exemplified in Figure 2c. In the case of the HMM taggers, where we begin with uniform estimates of both the state transition probabilities and the lexical probabilities, the learner finds it difficult to distinguish between more and less probable tag assignments.

We later discovered that previous implementations of UTBL involved limiting which possible part of speech assignments were placed into the lexicon[1], which was not explicitly detailed in the published reports. We then simulated, in a similar fashion, the construction of higher quality lexicons by using relative frequencies of tags for each word from the tagged Treebank to limit allowable word-tag assignments. That is, tags that appeared the tag of a particular word less than X% of the time were omitted from the set of possible tags for that word. We varied this threshold until accuracy did not significantly change on our set of heldout data. The effect of thresholding tags based on relative frequency in the training set is shown for our set of part-of-speech taggers in the curve in Figure 3. As shown in Table 1, the elimination of noisy possible part-of-speech assignments raised accuracy back into the realm of previously published results. The best test set accuracies for the learners in the class of HMM taggers are

[1] Eric Brill, Personal Communication

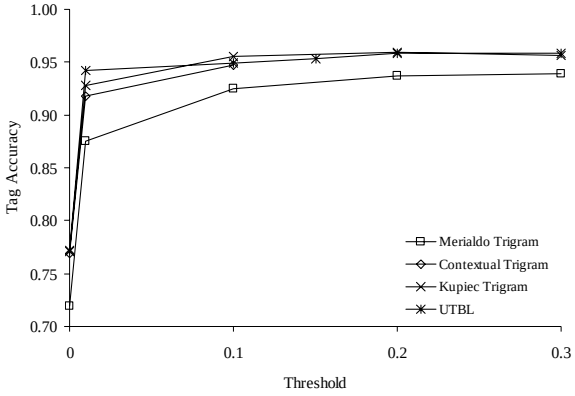

Figure 3: The effect of lexicon construction on unsupervised part-of-speech taggers

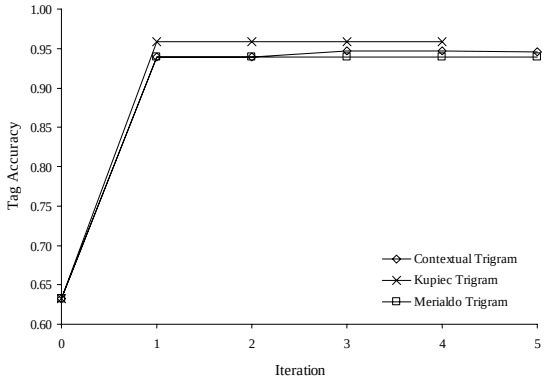

Figure 4: Test Accuracy of HMMs using Optimzed Lexicons

plotted against the number of training iterations in Figure 4.

5 Unsupervised Training With Noisy Lexicons

While placing informed limitations on the tags that can be included in a lexicon can dramatically improve results, it is dependent on some form of supervision – either from manually tagged data or by a human editor who post-filters an automatically constructed list. In the interest of being as unsupervised as possible, we sought to find a way to cope with the noisy aspects of the unfiltered lexicon described in the previous section.

We suspected that in order to better control the training of lexical probabilities, having a stable model of state transition probabilities would be of help. We stabilized this model in two ways.

	Unfiltered Lexicon	Optimized Lexicon
Merialdo HMM	71.9	93.9
Contextualized HMM	76.9	94.0
Kupiec HMM	77.1	95.9
UTBL	77.2	95.9
Contextualized HMM with Classes	77.2	95.9

Table 1: Tag Accuracy of Unsupervised POS Taggers

5.1 Using Unambiguous Tag Sequences To Initialize Contextual Probabilities

First, we used our unfiltered lexicon along with our tagged corpus to extract non-ambiguous tag sequences. Specifically, we looked for trigrams in which all words contained at most one possible part-of-speech tag. We then used these n-grams and their counts to bias the initial estimates of state transitions in the HMM taggers. This approach is similar to that described in Ratnaparhki (1998), who used unambiguous phrasal attachments to train an unsupervised prepositional phrase attachment model.

5.2 HMM Model Training Revised

Second, we revised the training paradigm for HMMs, in which lexical and transition probabilities are typically estimated simultaneously. We decided to train the transition model probabilities first, keeping the lexical probabilities constant and uniform. Using the estimates initially biased by the method previously mentioned, we train the transition model until it reaches convergence on a heldout set. We then use this model, keeping it fixed, to train the lexical probabilities, until they eventually converge on heldout data.

5.3 Results

We implemented this technique for the Kupiec, Merialdo and Contextualized HMM taggers. From our training data, we were able to extract data for on the order of 10,000 unique unambiguous tag sequences which were then be used for better initializing the state transition probabilities. As shown in Table 2, this method improved tagging accuracy of the Merialdo and contextual taggers over traditional simultaneous HMM training, reducing error by 0.4 in the case of Merialdo and 0.7 for the contextual HMM part-of-speech tagger.

HMM Tagger	Simultaneous Model Training	Sequential Model Training
Merialdo	93.9	94.3
Contextualized	94.0	94.7
Kupiec	95.9	95.9

Table 2: Effects of HMM Training on Tagger Accuracy

In this paradigm, tagging accuracy of the Kupiec HMM did not change.

6 Contextualized Tagging with Supervision

As one more way to assess the potential benefit from using left and right context in an HMM tagger, we tested our tagging model in the supervised framework, using the same sections of the Treebank previously allocated for unsupervised training, development and testing. In addition to comparing against a baseline tagger, which always chooses a word's most frequent tag, we implemented and trained a version of a standard HMM trigram tagger. For further comparison, we evaluated these part of speech taggers against Toutanova et al's supervised dependency-network based tagger, which currently achieves the highest accuracy on this dataset to date. The best result for this tagger, at 97.24%, makes use of both lexical and tag features coming from the left and right sides of the target. We also chose to examine this tagger's results when using only $<t_i, t_{i-1}, t_{i+1}>$ as feature templates, which represents the same amount of context built into our contextualized tagger.

As shown in Table 3, incorporating more context into an HMM when estimating lexical probabilities improved accuracy from 95.87% to 96.59%, relatively reducing error rate by 17.4%. With the contextualized tagger we witness a small improvement in accuracy over the current state of the art when using the same amount of context. It is important to note that this accuracy can be obtained without the intensive training required by Toutanova et. al's log-linear models. This result falls only slightly below the full-blown training-intensive dependency-based conditional model.

7 Conclusions

We have presented a comprehensive evaluation of several methods for unsupervised part-of-speech tagging, comparing several variations of hidden Markov model taggers and unsupervised transformation-based learning using the same corpus and same lexicons. We discovered that the quality of the lexicon made available to unsupervised learner made the greatest difference to tagging accuracy. Filtering the possible part-of-speech assignments contained in a basic lexicon automatically constructed from the commonly-used Penn Treebank improved results by as much as 22%. This finding highlights the importance of the need for clean dictionaries whether they are constructed by hand or automatically when we seek to be fully unsupervised.

Supervised Tagger	Test Accuracy
Baseline	92.19
Standard HMM	95.87
Contextualized HMM	96.59
Dependency Using LR tag features	96.55
Dependency Best Feature Set	97.24

Table 3: Comparison of Supervised Taggers

In addition, we presented a variation on HMM model training in which the tag sequence and lexical probabilities are estimated in sequence. This helped stabilize training when estimation of lexical probabilities can be noisy.

Finally, we experimented with using left and right context in the estimation of lexical probabilities, which we refer to as a contextualized HMM. Without supervision, this new HMM structure improved results slightly compared to a simple trigram tagger as described in Merialdo, which takes into account only the current tag in predicting the lexical item. With supervision, this model achieves state of the art results without the lengthy training procedure involved in other high-performing models. In the future, we will consider making an increase the context-size, which helped Toutanova et al. (2003).

8 Acknowledgements

The authors wish to thank Gideon Mann for performing some initial experiments with a publicly available implementation of UTBL, and Eric Brill for discussions regarding his work on unsupervised transformation based learning.

References

L.R. Bahl, F. Jelinek, and R. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179--190.

L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164-171.

E. Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, ACL*. Trento, Italy.

E. Brill. 1995. Unsupervised learning of disambiguation rules for part of speech tagging. *In Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA.

K. Church. 1998. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, ACL.

M. Collins. 2002. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA.

D. Cutting, J. Kupiec, J. Pedersen and P. Sibun. 1992. A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing*. ACL.

D. Elworthy. 1994. Does Baum-Welch re-estimation help taggers. In *Proceedings of the Fourth Conference on Applied Natural Language Processing, ACL*.

J. Kupiec. 1992. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language* 6.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML-01*, pages 282-289.

B. Merialdo. 1991. Tagging English text with a probabilistic model. In *Proceedings of ICASSP*. Toronto, pp. 809-812.

H. Ney, U. Essen and R. Kneser. 1994. On structuring probabilistic dependencies in stochastic language modeling. *Computer, Speech and Language*, 8:1-38.

A. Ratnaparkhi. 1998. Unsupervised statistical models for prepositional phrase attachment. In *Proceedings of the Seventeenth International Conference on Computational Linguistics*. Montreal, Canada.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL*. pp. 252-259.

A.J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260--269.

S. Young. 1999. Acoustic modelling for large vocabulary continuous speech recognition. *Computational Models of Speech Pattern Processing: Proc NATO Advance Study Institute*. K. Ponting, Springer-Verlag: 18-38.