Joint Labeling of Multiple Sequences: A Factorial HMM Approach

Kevin Duh
Department of Electrical Engineering
University of Washington, Seattle, USA
duh@ee.washington.edu

Abstract

Various sequence-labeling tasks in natural language processing require the cascading of error-prone subtasks. For instance, in syntactic analysis of sentences, part-of-speech (POS) tagging results are often used for noun-phrase segmentation (NP chunking), even though initial errors may hurt downstream processing. To mitigate this problem, I use a factorial hidden Markov Model (FHMM), where the POS and NP are contained in the two hidden state sequences. I examine whether information sharing between POS and NP sequences will improve performance in both subtasks. The results of FHMM is compared to a baseline of cascaded HMM POS tagger and NP chunker. Using a FHMM with switching-parent dependencies, a state-of-the-art NP chunking accuracy of 98.3% is achieved.

1 Introduction

Traditionally, various sequence labeling problems in natural language processing and bioinformatics are solved by the cascading of well-defined subtasks, each extracting specific knowledge. For instance, the problem of information extraction from sentences may be broken into several stages: first, part-of-speech (POS) tagging is performed on the sequence of word tokens; this result is then utilized in noun-phrase and verb-phrase segmentation. Finally, a higher-level analyzer extracts relevant information based on phrase segmentation and POS tagging knowledge.

The decomposition of problems into well-defined subtasks is useful but sometimes lead to errors. The problem with this approach is that errors in earlier subtasks will propagate to downstream subtasks, which deteriorates performance. Therefore, a method that allows *joint labeling* of subtasks is desired. Two major advantages arise from simultaneous labeling of subtasks: First, there is more robustness against error propagation. This is especially relevant if we use probabilities in our models. Cascading subtasks inherently "throws away" the probability at each stage; joint labeling does not. Second, information between simultaneous subtasks can be shared to further improve estimation.

Previous work on joint labeling include the use of N-best Rescoring (Xun et al., 2000), an extension of transformation-based learning (Florian and Ngai, 2001), and conditional random fields (McCallum et al., 2003). In this paper, I explore the use of factorial hidden Markov models (FHMM) for joint labeling of multiple sequences. To the best of my knowledge, this is the first application of FHMM to joint POS-NP labeling. The paper is structured as follows: Section 2 discusses the FHMM and its extensions. Section 3 explains the POS tagging and NP chunking subtasks, and the data used in experiments. Section 4 reports experimental results and Section 5 concludes.

2 Factorial HMM

A FHMM is an HMM with distributed state representation. Ghahramani first developed it based on the motivation that real world observations are often caused by the interaction of multiple independent causes (Ghahramani and Jordan, 1997). It has been applied much in source separation problems (e.g. speaker separation (Deoras and Hasegawa-Johnson, 2004)) and is shown to model loosely-coupled sequences effectively. Here, I will use FHMMs to model the loosely-coupled sequences of POS tags and NP labels.

2.1 Basic FHMM

The basic FHMM graph is shown in Figure 1. From it we can infer several conditional independence statements, such as: the states X_t and Y_t are conditionally dependent given the observation Z_t. The joint probability of the state

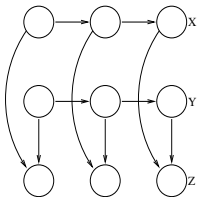


Figure 1: Basic FHMM Graph

sequences and observations is given by:

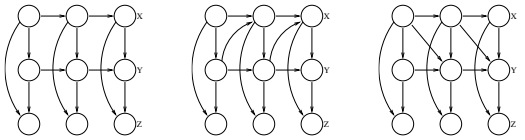
$$p(x_{1:T}, y_{1:T}, z_{1:T}) \quad (1)$$
$$= \prod_{t=1}^{T} p(x_t|x_{t-1})p(y_t|y_{t-1})p(z_t|x_t, y_t)$$

where $p(x_1|x_0) = \pi(x_1)$, etc. represent the initial distributions.

The FHMM is topologically equivalent to a HMM with states being the cross-product of individual FHMM states. Despite of this equivalence, FHMM is advantageous because, if correctly modeled, requires the estimation of substantially fewer parameters.

2.2 Additional FHMM Structures

Many other structures exist in the FHMM framework. Statistical modeling often involves the iterative process of constraining and relaxing independence statements. Additional edges may be added across states in the graphical model to represent the dependencies present in the data. Figures 2(a), 2(b), and 2(c) show three such models, which will be used for the experiments described later.



(a) (b) (c)

Figure 2: Graphs of FHMM with additional dependencies.

In each of these graphs, the top, middle, and bottom state sequences represent the NP labels, POS tags, and observed word variables, respectively. In FHMM 2(a), edges are added between NP and POS since it is conceivable that knowledge of one significantly affect the estimate of the other. For instance, if we know the current word is in a noun-phrase, then the probability of the POS tag being a noun will very likely increase.

Independence statements may also be relaxed by the use of switching parents (Bilmes, 2000). Switching parents allow a child variable to change its parents depending on the value of another variable, as shown in Figure 3. In Figure 4(a), the POS tag is connected by either its previous POS tag or NP tag, depending on the value of variable S, which in turn depends on the previous POS tag and current NP tag. This is a form of discriminative structure learning, which picks the situations where NP or POS tags are more helpful as conditional parents. Figure 4(b) is more elaborate switching-parent FHMM, which switches between the models in Figures 2(b) and 2(a). This is useful when we would like to add the dependency between the previous POS tag and current NP tag, but sometimes suffer from spare data problems.

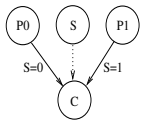
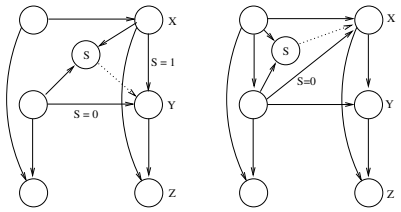


Figure 3: Switching Parents. Parent(C) is either P0 or P1 depending on the value of S



(a) (b)

Figure 4: Switching Parent FHMMs.

3 POS Tagging and NP Chunking

3.1 The Tasks

POS tagging is the task of assigning words the correct part-of-speech, and is often the first stage of various natural language processing tasks. As a result, POS tagging has been one of the most active areas of research, and many statistical and rule-based approach have been tried. The most notable of these include the trigram HMM tagger (Brants, 2000), maximum entropy tagger (Ratnaparkhi, 1996), and transformation-based tagger (Brill, 1995).

Accuracy numbers for POS tagging are often reported in the range of 95% to 97%. Although this may seem high, note that a tagger with 97% accuracy has only a 63% chance of getting all tags in a 15-word sentence correct, whereas a 98% accurate tagger has 74% (Manning and Schütze, 1999). Therefore, small improvements can be significant, especially if downstream processing requires correctly-tagged sentences. One of the most difficult problems with POS tagging is the handling of out-of-vocabulary words.

Noun-phrase chunking is the task of finding the non-recursive (base) noun-phrases of sentences. This segmentation task can be achieved by assigning words in a sentence to one of three tokens: `BEGIN-NP`, `INSIDE-NP`, or `OUTSIDE-NP` (Ramshaw and Marcus, 1995). The state-of-the-art chunkers report 93%-94% F1 measure and accuracies of 87%-97%. (See, for example, NP chunkers utilizing conditional random fields (Sha and Pereira, 2003) and support vector machines (Kudo and Matsumoto, 2001).)

3.2 Data

The data comes from the CoNLL 2000 shared task (Sang and Buchholz, 2000), which consists of sentences from the Penn Treebank Wall Street Journal corpus (Marcus et al., 1993). The training set contains a total of 8936 sentences with 19k unique vocabulary. The test set contains 2012 sentences and 8k vocabulary. Of the 8k vocabulary, 2.5k are out-of-vocabulary words. On the other hand, in terms of raw word token count, there are 7% out-of-vocabulary words in the test set.

There are 45 different POS tags and 3 different NP labels in the original data. To reduce inference time, I collapsed the 45 POS tags to 5 labels: `NOUN`, `VERB`, `ADJECTIVE`, `ADVERB`, and `OTHER`. This procedure is similar to (McCallum et al., 2003); therefore our results can be compared. An example sentence with POS and NP tags is shown in Table 1.

The	move	could	pose	a	challenge
ADJ	NN	V	V	ADJ	NN
I	I	O	O	I	I

Table 1: Example sentence with POS tags (2nd row) and NP labels (3rd row). For NP, I = Inside-NP, O=Outside-NP.

4 Experiments

Two sets of experimental results will be reported here. For all models, parameter estimation is performed using Witten-Bell smoothing, since initial experiments show it has the best results. The first set of experiments compare the performance of cascaded HMMs vs. various FHMM structures. The POS and NP labeling accuracy is shown in Table 4. Note that an HMM NP Chunker operating on correct POS tags achieves 87.9% NP accuracy, whereas the same chunker operating on the results of a 93.8% accurate HMM POS tagger (in cascaded fashion) achieves only 48.7% NP accuracy. I conjecture that the extremely low NP accuracy is due to the compounding effects of having a bigram NP chunker with a low accuracy POS tagger.

Note that the performance of FHMM 2(a), 2(b), and 2(c) exceed that of the HMM NP chunker with perfect POS observations, which suggests that information sharing between NP and POS tags play an important role in improved labeling. More specifically, Table 3 shows the POS accuracy for Out-of-Vocabulary (OOV) words only. OOV words are usually the root cause of a POS taggers errors. Compared to the HMM and FHMM 1, the latter three FHMM models perform much better at guessing the POS tag of unknown words. This implies that for OOV words, an additional edge between NP and POS states is extremely beneficial.

Model	POS	NP
HMM for NP only	100	87.9
Cascaded HMM	93.8	48.7
FHMM 1	93.2	87.9
FHMM 2(a)	**95.2**	93.0
FHMM 2(b)	94.9	**93.7**
FHMM 2(c)	95.1	93.5

Table 2: POS and NP Accuracy for Cascaded HMM and FHMM Models. The number identifying each FHMM corresponds to the figure label.

Table 4 shows the second set of results obtained with switching parents FHMMs. The numbers are intriguing. The poor results obtained for FHMM 4(a) is expected, since it switches off essential dependency links (e.g. POS state-to-state transition). On the other hand, FHMM 4(b) shows surprisingly high NP accuracy and equally surprisingly low POS ac-

Model	Overall	%OOV-error	OOV
HMM	93.7	67.9	39.1
FHMM 1	93.2	62.4	38.7
FHMM 2(a)	95.2	61.0	57.7
FHMM 2(b)	94.9	64.1	53.2
FHMM 2(c)	95.1	60.1	57.9

Table 3: Relative Effect of OOV words on POS Accuracy. The columns represent Overall Accuracy (%), Percentage of errors that are OOV words, and Accuracy for OOV words only.

curacy. I first designed FHMM 4(b) as a middle-ground between FHMM 2(a) and FHMM 2(b), since each individually has the highest respective POS and NP accuracy thus far. I conjectured initially that the switching parent version will strike a good balance between the two. Instead, it seemed that it has optimized for NP accuracy at the expense of POS accuracy. Looking at the probability table of the switching parent reveal that the additional edge is added (S=0) usually when the following situations regarding the previous NP and POS tags occur: (1) NOUN, BEGIN-NP; (2) VERB, BEGIN-NP; (3) VERB, IN-NP; (4) ADVERB, IN-NP. I have not yet found an linguistic explanation.

Model	POS	NP
FHMM 2(a)	95.2	93.0
FHMM 2(b)	94.9	93.7
FHMM 4(a)	81.5	71.4
FHMM 4(b)	53.1	**98.3**
DCRF (McCallum '03)	82.0	86.1
Transformation (Florian '01)	**96.6**	97.2

Table 4: Comparison of POS and NP Accuracy for Switching Parents FHMM and techniques from other papers.

5 Conclusion

The problem of joint POS and NP labeling is addressed using various FHMM structures. Results show that information sharing between POS and NP tags significantly improve both POS and NP accuracy. A FHMM with switching parents achieved a NP accuracy of 98.3%, which is better than state-of-the-art results in the literature. This work has shown that FHMM with switching parents is a promising framework for joint labeling.

Acknowledgements

I would like to thank Katrin Kirchhoff, Jeff Bilmes, Chris Bartels, Xiao Li, and Gang Ji for inspiring discussions on this work and help on GMTK software.

References

Jeff Bilmes. 2000. Dynamic bayesian multinetworks. In *The 16th Conference on Uncertainty in Artificial Intelligence*, Stanford.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)*.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–565.

A. N. Deoras and Mark Hasegawa-Johnson. 2004. A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel. In *ICASSP 2004*.

Radu Florian and Grace Ngai. 2001. Multidimensional transformation-based learning. In *Proceedings of the 5th Computational Natural Language Learning Workshop (CoNNL-2001)*.

Z. Ghahramani and M. I. Jordan. 1997. Factorial hidden Markov models. *Machine Learning*, 29:245–275.

T. Kudo and Y. Matsumoto. 2001. Chunking with support vector machines. In *Proceedings of NAACL-2001*.

C. D. Manning and H. Schütze, 1999. *Foundations of Statistical Natural Language Processing*, chapter 10. MIT Press.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Andrew McCallum, Khashayar Rohanimanesh, and Charles Sutton. 2003. Dynamic conditional random fields for jointly labeling multiple sequences. In *NIPS 2003 Workshop on Syntax, Semantics, Statistics*.

L. A. Ramshaw and M. P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora (ACL-95)*.

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

E. F. Tjong Kim Sang and S. Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. In *Proceedings of CoNLL-2000*, pages 127–132.

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL*.

E. Xun, C. Huang, and M. Zhou. 2000. A unified statistical model for the identification of English BaseNP. In *Proceedings of ACL*, pages 109–116.