Hidden Markov Models vs. Maximum Entropy Markov Models

Ivan Oropeza
Dept. of Computer Science
University of Texas at Austin

William Xie
Dept. of Computer Science
University of Texas at Austin

1. INTRODUCTION

A frequent problem in many disciplines is the challenge to do sequence labelling. DNA sequencing, video semantic analysis, and Part-Of-Speech tagging are just some examples where sequence labelling is the underlying problem that needs to be solved [2, 6, 4]. In Natural Language Processing (NLP), Part-Of-Speech (POS) or lexical category tagging is an important problem because it serves as a stepping stone into solving more complex problems such as semantic analysis of sentences. The typical techniques applied to this problem are Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM), and Conditional Random Fields (CRF). While CRFs are considered to be the state-of-the-art in POS tagging we want to compare the performance of the other models, HMM and MEMM. This paper is structured in the following way: first we provide a brief description of the datasets used in this investigation. Then, we discuss some of the differences between HMMs and MEMMs. Next, we explain how the training process works specifically for POS tagging. Finally, we compare the performance of HMMs vs MEMMs under similar circumstances.

2. DATASETS

In this investigation we use two corpora. The first one is the Wall Street Journal (WSJ) corpus. The WSJ corpus is a collection of 2,499 articles collected in a span of three years from the Wall Street Journal. It has approximately 3 million words and was tagged by using statistically-based methods. This corpus has a total of 82 possible tags citewsjCorpus. Its range of topics is very narrow and thus it is reasonable to expect that the word choice distribution in the articles to be narrower than other corpora as well. Hence, we expect better performance when we test with this metric.

In addition to the WSJ corpus, we also use the Brown corpus. The Brown corpus is a manually tagged collection of 500 text documents sampled from 1961. It uses 36 POS tags and it is consolidated from various sources and from various topics such as fiction, press, and lore [3]. Consequently, we expect that this corpus would represent a more general distribution on word choice than WSJ corpus. Therefore, we expect scores to be slightly lower than when using WSJ.

```
\[ He/PRP \]
tried/VBD to/TO ignore/VB
\[ what/WP \]

\[ his/PRP\$ own/JJ common/JJ sense/NN \]
told/VBD
\[ him/PRP \]
,/, but/CC
\[ it/PRP \]

\[ was/VBD n't/RB possible/JJ \]
;/: ;/:
\[ her/PRP$ motives/NNS \]
were/VBD too/RB blatant/JJ ./.
```

Figure 1: Example sentence from the Brown Corpus [3]

3. HMM VS. MEM,

A HMM is generative model for the joint distribution of states and observations. It follows a Markovian assumptions. The Markov assumption restricts the transitions between states to be dependent only on the immediate past [5]. On the other hand, MEMM is discriminative since it models the conditional probability of the state given the observation.

An MEMM is an enhancement on the Maximum Entropy model, also known as a multinomial regression model which in turn attempts to do classification by making the fewest number of assumptions. Figure 2 shows the pictorial differences between both graphical models. The added benefit of MEMMs over HMMs is that MEMMs are not limited to only modelling two aspects: $P(S_i|P_{i-1})$ and $P(O_i|S_i)$. Instead, MEMMs consider information derived directly from features applied on the observation in addition to the previous state knowledge. For example, capitalization often occurs with Nouns and specific suffixes such as ”ed” and ”ing” tend to be associated with Verbs are some useful features that can not be modelled by an HMM but can be modelled by an

MEMM. Moreover, the typical algorithm used to solve the decoding problem in an HMM can be trivially modified to solve the decoding problem in an MEMM without additional overhead [5].

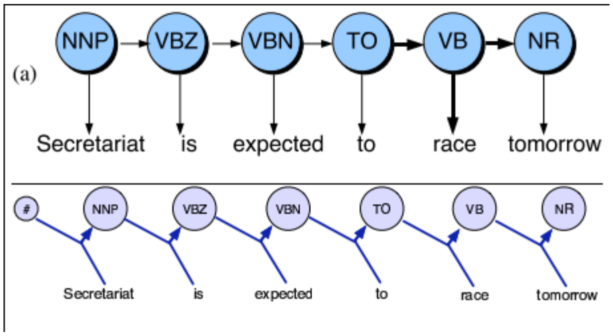

Figure 2: Hidden Markov Model(top) and Maximum Entropy Markov Model(bottom) [5]

4. LEARNING

In order to train our HMM model in a supervised manner

5. LAPLACE SMOOTHING

6. EXPERIMENT

Corpus	**Train/Test**	$\overline{Sentences}$	$\sigma Sentences$	$\overline{Tags}$	$\sigma Tags$

Figure 3: HMM scores for the Brown and WSJ corpora. The first column is the corpus used. The second column is the data used for evaluating the system. Testing with training data will give us an idea of overfitting and testing with testing data will give us an idea of effectiveness of the system. Four scores are reported: the average and stdev for the number of sentences correctly tagged and the average and stdev for the number of tags correctly labelled. The experiments

Corpus	**Train/Test**	$\overline{Sentences}$	$\sigma Sentences$	$\overline{Tags}$	$\sigma Tags$

Figure 4: MEMM scores for the Brown and WSJ corpora. The first column is the corpus used. The second column is the data used for evaluating the system. Testing with training data will give us an idea of overfitting and testing with testing data will give us an idea of effectiveness of the system. Four scores are reported: the average and stdev for the number of sentences correctly tagged and the average and stdev for the number of tags correctly labelled. The experiments

7. REFERENCES

[1] E. Charniak. BLLIP 1987-89 WSJ Corpus Release 1, 2000.

[2] W. Ching, E. Fung, and M. Ng. Higher-order hidden markov models with applications to dna sequences. In J. Liu, Y.-m. Cheung, and H. Yin, editors, *Intelligent Data Engineering and Automated Learning*, volume 2690 of *Lecture Notes in Computer Science*, pages 535–539. Springer Berlin Heidelberg, 2003.

[3] W. N. Francis and H. Kucera. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[4] F. Hasan, N. UzZaman, and M. Khan. Comparison of different pos tagging techniques (n-gram, hmm and brillâĂŹs tagger) for bangla. In K. Elleithy, editor, *Advances and Innovations in Systems, Computing Sciences and Software Engineering*, pages 121–126. Springer Netherlands, 2007.

[5] D. Jurafsky and J. H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[6] G. Xu, Y.-F. Ma, H.-J. Zhang, and S.-Q. Yang. An hmm-based framework for video semantic analysis. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(11):1422–1433, Nov 2005.