**CMPT 825: Natural Language Processing** **Spring 2008**

# Lecture 8 — Jan 23, 2008

*Lecturer: Anoop Sarkar* *Scribe: Anton Venema*

The paper under consideration for this scribing report is *A Second-Order Hidden Markov Model for Part-of-Speech Tagging* by Scott M. Thede and Mary P. Harper. In it, they describe a new approach for using HMMs in part-of-speech tagging. Using existing standards for contextual information, they add more detailed lexical information and tag prediction for unknown words using suffix data. There are three key components to their model:

1. A standard contextual trigram tagger with probabilities: $P(t_i|t_{i-1}, t_{i-2})$
2. A bigram lexical tagger with probabilities: $P(w_i|t_i, t_{i-1})$
3. An bigram unknown-word tagger with probabilities: $P(s_i|t_i, t_{i-1})$

The notation used for the probability descriptions was somewhat confusing. I believe the descriptions above say the same thing with more understandable notation. The lexical and unknown-word taggers were at first going to be trigram taggers, but the results were unsatisfactory. Since examining the current and previous two tags proved to be detrimental, only the current and previous one tag was considered in the final results. The Jelinek-Mercer method of smoothing (Jelinek and Mercer, 1980) was used to help reduce sparseness in the data The three equations below represent the contextual, lexical, and unknown data probability distributions respectively.

$$\hat{P} = k_3 \cdot \frac{N_3}{C_2} + (1 - k_3)k_2\frac{N_2}{C_1} \cdot (1 - k_3)(1 - k_2) \cdot \frac{N_1}{C_0} \tag{8.1}$$

$$\hat{P} = (\frac{log(N_3 + 1) + 1}{log(N_3 + 1) + 2})\frac{N_3}{C_2} + (\frac{1}{log(N_3 + 1) + 2})\frac{N_2}{C_1} \tag{8.2}$$

$$\hat{P}_i = f(N_i)\hat{c}_i(s_i) + (1 - f(N_k))\hat{P}_i(s_i - 1), 1 < k \leq 4 \tag{8.3}$$

The unknown data probability is recursive, with $\hat{P}_i = \hat{c}_1, k = 1$ providing the base case. (The subscripts have been rewritten to improve readability.) This method gives more weight to longer and more frequently appearing

suffixes. An alternative smoothing method was proposed that would use word class information (see Taoukermann and Radev, 1996).

The results indicated an improvement over standard bigram/trigram taggers and over HMMs using only second-order lexical probabilities. The comparisons to other researchers was a bit weak, since the alternatives available for closed lexicon used different training and test data. In spite of this, the results show a high improvement over alternative training methods.

I would be interested to see if these results hold up in languages other than English - specifically a language which does not have a high degree of suffixation. I would also be interested to see if any improvements would result from a model that doesn't discriminate based on suffixation only, but prefixation as well.