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

Lecture 9 — Jan 25, 2008

Lecturer: Anoop Sarkar *Scribe: Anton Venema*

The paper under consideration for this scribing report is *Does Baum-Welch Re-estimation Help Taggers?* by David Elworthy. It yields results that are strikingly consistent with those demonstrated by Bernard Merialdo (1994), who concluded that the best way to get a high-performance model was to use as much tagged data as possible, and only use Baum-Welch re-estimation (using the Forward-Backward algorithm) if the amount of tagged data available is minimal. He additionally concluded that even when Baum-Welch re-estimation improves model accuracy, the number of iterations used should be kept minimal, since multiple iterations tend to degrade the performance in the long run.

Elworthy had the goal in mind not of refuting Merialdo's results, but providing more detailed information as to why he could draw the conclusions he did. His end results were almost identical. When using HMMs, he recommends using as much tagged text as possible to train the model. If the test data is similar to the training data, then BW re-estimation should be used minimally, if at all. For less robust starting conditions (i.e. limited corpora), use as much pre-conditioned information as possible, and then use BW re-estimation minimally. If no training data is available, use BW re-estimation, but only up to a point where it ceases to be beneficial (i.e. iteration x_i shows no improvement over iteration x_{i-1}). It is important to note that the effectiveness of BW re-estimation even at this point is highly dependent on good initial lexical (emission) and transition probabilities.

While the information presented by Elworthy is more detailed than that provided by Merialdo, one has to question its necessity. Very little new information was presented, and no new conclusions were drawn. In fact, Elworthy concludes his paper by restating Merialdo's conclusions. The new information that Elworthy gives is the classification of models into one of three categories - initial, early, and classical - based on their responsiveness to BW re-estimation. Even so, this seems to merely be giving a name to what Merialdo previously reported.

- Classical - rising accuracy on each iteration

- Initial maximum - highest accuracy at outset, degradation at each iteration

- Early maximum - rising accuracy on first few iterations, then degradation

Classical patterns can be observed with no initial training data. Models exhibiting early maximum patterns can be observed when their initial probability distributions were obtained through a supervised learning process on a relatively small (¡ 100,000 sentences) amount of training data. Initial maximum patterns can be seen for models attained in much the same way, but with a larger amount of training data.