Corpus Navigator Status

Jonathan Clark

Monday, February 11, 2008

1 TODO (Old Business)

1. Schedule meeting with Vamshi, Alon, and Jaime (Alon out of town soon)
2. Ask linguistics grad student to fill out spreadsheet and glosses for feature detection gold standard
3. LREC Notification Tomorrow: If accepted, I owe them about 50 WALS feaures by March 24.

2 Upcoming Experiment Questions

1. Will the gold standard for navigation be learning which features are grammaticalized using the least number of sentences? If so, I think navigation will either have to select less than 3000 sentences, or we will have to elicit more sentences from an Urdu speaker.
2. What metric will we use for evaluation? f-measure? If so, does this imply we have to know what deductions can be made from each set of sentences?

3 Feature Detection Questions

1. "Lexical clustering" is still creating difficulties in feature detection. Have I missed something? (Note: weighting evidence based on number of changed words will help this problem somewhat, I think, but this is a matter of both tractability and accuracy).

4 Current Issues (Old Business)

1. Still haven't cleaned up all of elicitation corpus. Some feature structures remain corrupt. Unsure of how much time will be required to fix this.
2. Working on filtering noise from feature detection (almost done coding weights based on number of changed words)

5 NIST Experiment Status

1. Since Vamshi saw no improvement from adding 90,000 rules to the system, the experiment with changing 100 input sentences is on hold
2. No word from Philip or Chris

6 Recently Accomplished

1. Optimizations in feature detection code (about 3X speedup)