

# Corpus Navigator Status

Jonathan Clark

Monday, August 18, 2008

## 1 What's New

1. Spent most of my week improving lexical clustering and attempting to track down the source of the score decrease found last week.
2. Implemented use of head matching as means of lexical clustering (helped)
3. Question from last week: Why did scores go down in response to a cleaner corpus? They didn't: The cleaner corpus helped scores, but some change in the code (which I'm still searching for) appears to have changed the results.
4. Is my method for following and projecting dependencies reasonable?

Results using word alignment and phi features:

|                   | f1    | prec  |        | recall |        |   |
|-------------------|-------|-------|--------|--------|--------|---|
| baseline          | 0.218 | 0.162 | 35/215 | 0.333  | 35/105 | everything                                    |
| only me, dep, gov | 0.186 | 0.131 | 34/259 | 0.323  | 34/105 | no other                                      |
| only me           | 0.179 | 0.128 | 31/241 | 0.295  | 31/105 | only me                                       |
| baseline (asim)   | 0.152 | 0.230 | 12/52  | 0.114  | 12/105 | everything (asim's corrections)               |
| better lex clust  | 0.211 | 0.238 | 20/84  | 0.190  | 20/105 | (added "backoff head" for "be" verbs) <- prec |
| lower thresholds  | 0.211 | 0.180 | 27/150 | 0.257  | 27/105 | (theme threshold down to .01) <- recall       |

Results of overlap with baseline: 41/88

## 2 Todo

1. Error analysis of feature detection to find method of increasing precision (on-going)
2. Turn notes into Opening Book Corpus (Tues)
3. Finish heuristic scoring mechanisms (Wed-Sat)
4. Finish gathering heuristic knowledge from Lori (need to arrange meetings for this week and next week)
5. Test Corpus Navigator (next week)
6. Experiments: Factored MT and Avenue MT