Corpus Navigator Status

Jonathan Clark

Monday, August 4, 2008

1 What's New

1. Since word alignments and head mappings are not easily human readable, I didn't get very far in a manual evaluation of how word alignments would help
2. Implemented use of word alignments and me, my dependent, and my governor in feature detection
3. 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

Results of overlap with baseline: 41/88
```

2 Todo

1. Error analysis of feature detection
2. Turn notes into Opening Book Corpus
3. Write heuristic evaluation code (2/5 modules done)
4. Rewrite heuristic elicitor (DONE)
5. Finish gathering heuristic knowledge from Lori
6. Test Corpus Navigator
7. Write "marked-on" code (DONE)
8. Run factored MT experiment
9. Run Avenue MT experiment
10. As time permits, implement decision tree type model for increasing FD precision

3 Open Research Issues (No Updates)

1. Even our heuristic currently uses manually tuned weights, which people will probably complain about. If we want to use machine learning on them, we need some sort of training data.
2. Can we simulated factored translation in Avenue? Or do we have to use feature constraints to get the same benefits?
3. What lexical resources do we plan to use for our Avenue experiments?
4. How do we prevent overfitting? (Should I use machine learning-type optimization anyway?)
5. We don't consider probability mass of language covered