Corpus Navigator Status

Jonathan Clark

Monday, June 30, 2008

1 What's New

1. Need to rewrite knowledge elicitor to allow more expressive formalism
2. Morphological Analyzer *theoretically* works, but I'm currently ironing out some disk read/write issues

2 Project Issues (No Updates)

1. Need to schedule time for continuing to build knowlegebases.
2. Implementation of a non-interactive interface with ParaMor is complete. Now integrating his code directly into feature detection.
3. As soon as I have this and the morph analyzer finished, I will begin constructing a factored Moses system by using only the correct morphemes since I wasn't able to fit the precision filter into this summer's schedule.
4. Implementation of navigation "simulator" that ranks sentences with no knowledge source is behind, but I'm putting priority on tasks that require collaboration with Christian since he'll be gone soon.
5. Still need to turn general guidelines from our previous meeting into an actual Opening Book.

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

4 Research Plan

Primary Goal: Get all pieces together so we have a deliverable.
Intermediate Goal: Show that navigation can affect rule learning which can affect translation quality (perhaps Paul can help).
Intermediate Goal: Compare 1-best translations to oracle-best and see if feature constraints could help.

1. Create navigation "simulator" that simulates navigation without interaction from a real user (all scoring functions and knowledge sources are dummy stubs, but we can request the actions: translate, align new, or align with suggestion) *By SAT 5/24 – **70% PROGRESS***
2. Get Finsihed Urdu Gold Standard from Paul *By 6/2 – **COMPLETE***
3. Decide Opening Book in Long Meeting *By MON 6/9 – **50% PROGRESS***
4. Build Java API around ParaMor *By MON 6/9 – **COMPLETE***
5. Integrate ParaMor (add external API to ParaMor and associate our morphemes with Christian's paradigms) *By MON 6/9 – **70% PROGRESS***
6. Build an external morphological analyzer that segments with ParaMor and then assigns our features *By FRI 6/13 – **70% PROGRESS***
7. Implement heuristic evaluator with lookahead *By MON 6/30 – **70% PROGRESS***
8. Build MILES Implicational Universal Knoweldgebase ("If you know there's animacy marking, what else do you expect to be marked?") ***30% PROGRESS***
9. Build MILES "Related Features" Knowledgebase ("If you know there's animacy marking, what do you expect might interact with animacy?") ***30% PROGRESS***
10. Build WALS-to-MILES Feature Mapper *By FRI 7/4*
11. EXPERIMENT: Random Selection of Sentences vs Navigation of sentences (How many new feature constraints / morpheme-annotation pairings do we learn after each sentence?) – This shows that navigation can learn (almost) as much from less data *By FRI 7/4*
12. Prepare Urdu Parallel Training and Testing Data (includes running segmented and non-segmented data through training) *By FRI 7/11*
13. EXPERIMENT: Moses with and without our annotated morphology "factors" – This shows the data from Feature Detection is useful to SMT *By MON 7/14*
14. Implement Feature Constraint Generator *By MON 7/28*
15. Axed? EXPERIMENT: Avenue Rule Learner trained with and without our annotated morphology factors (both experiments use external lexical resources) – This shows Segmentation is useful to XFER? – Could help rule learning associate morphemes.
16. EXPERIMENT: Avenue Rule Learner trained on whole EC and then selected portion of EC (includes our annotated morphology factors AND our suggestions for feature constraints; both experiments use external lexical resources) – This shows Feature Detection is useful to Statistical XFER while using less data *By FRI 8/8*
17. Experiment with Precision Filtering *As time allows*