

## Individual Contributions

Arvind

For this project, I was primarily responsible for setting up the development environment, cleaning the data and generating datasets, and running the stemming, heldout, topic clustering, and nocorrect experiments. To do this, I wrote various scripts that sliced the corpus in different ways, including but not limited to:

- 1) identifying and dropping references and unwanted noise
- 2) correcting mislabeled annotations (some token boundaries were wrong in the dataset)
- 2) aligning the M2 gold file to any arbitrary system input
- 3) caching annotations using Redis to generate datasets with different error types
- 4) parsing the SGML into multiple documents to then be clustered and split in different ways
- 5) wrapping the M2 scorer into an easier to use `align_and_score` function call

To do stemming, I used the NLTK python module (<http://nltk.org/>), to do clustering, I wrote a few scripts that built on the patterns python module (<https://github.com/clips/pattern/>), and for topic analysis, I forked and modified an online LDA library (<https://www.github.com/arvs/streamLDA>) so that it would be low memory and deal with larger datasets (again, using Redis as a caching backend).

On the SMT side, I installed Moses, SRILM, IRSTLM, Docent, and a few other SMT tools on a friend's server, and did the necessary sysadmin work to manage multiple moses projects in parallel.

Louis

For my part, I did the initial research and writeup, and researched/experimented/wrote about the downsampling and significance testing approaches. I configured Moses to use SALM (<http://projectile.sv.cmu.edu/research/public/tools/salm/salm.htm#update>) and do significance testing (<http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc19>), and wrote scripts to handle automated significance testing for each system. I also wrote the scripts that downsampled each phrase table at different rates, and performed decoding and evaluation experiments on each of the systems Arvind created for different rates of downsampling and significance testing.