11-731: Machine Translation

Homework Assignment #10:
Out: Wednesday, April 15, 2009
Due: 2:00 pm Wednesday, April 29, 2009

Putting it all together: a real MT evaluation exercise

In previous homework assignments you have used a small Spanish-English training corpus which consists of 10K sentence pairs. In this homework you will use a considerably larger training corpus to build a full machine translation system. The data is provided at the following location: /afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW10. The data is already preprocessed. If you want to run your own preprocessing (for Part2), the original files are also available in ./original.

Part 1

Build a baseline machine translation system using the provided training data (i.e. training the alignment models, lexicalized distortion model, phrase table extraction). Use the English side of the training data to train a 3-gram language model. Run minimum error-rate training on the development set (./dev/dev_full.es). Use the optimized parameters to translate an unseen test set (./test/test.es) for which you don’t have the reference translations.

- Make reasonable assumptions as to which models to be used based on your experience in previous homework assignments. State your assumptions.
- Provide corpus statistics (e.g. corpus size, vocabulary size, average sentence length, etc) for the training data and test data (i.e. dev and test sets). Also the number of words (types and tokens) in the test data, which are not in the training data.
- Report results for the development set in BLEU, METEOR and TER metrics. Also provide an analysis on the coverage of the dev set. i.e. number of untranslated word (types and tokens), etc.
- You will be provided with the reference translations for the unseen test set one day prior to the submission deadline. Report the results for the test set in above 3 metrics and also provide an analysis on the coverage of the test set.

Part 2

Explore approaches to improve over the results of the baseline system. You may try (among other things):

- Better data processing approaches (e.g. splitting pronouns)
- Additional language model data
- Additional parallel data
- Using dictionaries
- Additional features in the phrase table
Report the results for both the development set and test set. Show improvements for each addition (i.e. if you have two additional approaches A1 and A2, your results table should give scores for the baseline, baseline+A1, baseline+A2, baseline+A1+A2).

Check if the improvements produce statistically significant results.

Bonus points will be awarded for creative ideas for improvement.

You are encouraged to look in to the proceedings of the Workshop on Machine Translation (WMT) conferences in 2008 and 2009 for ideas on improving the baseline system.