Assignment: Language Modeling

Professor: José Adrián Rodríguez Fonollosa

Objective: hands-on study of the performance of different strategies for language modeling, the trade-off between model order/size and perplexity reduction and the influence of the type and size of the training corpus.

The test corpus is the English part of the test set of the WMT2015 SMT evaluation.
http://www.statmt.org/wmt15/test.tgz

The recommended toolkit is the SRI LM toolkit

http://www.speech.sri.com/projects/srilm/

The training corpus can freely selected from the monolingual corpus as well as the English part of parallel corpus referenced in http://www.statmt.org/wmt16/translation-task.html

Task description:

- Download and install the SRI Language Modeling Toolkit

- Download training English corpus

- Obtain different language models with different methods and orders. You can also try different concatenations of corpus or corpus interpolation with the tool **compute-best-mix** http://www-speech.sri.com/projects/srilm/manpages/ppl-scripts.1.html

- Evaluate you language models with the ngram program and the test corpus (English sentences of the test set mentioned above)

- Show your results in a table and a graph