

11-731: Machine Translation

Homework Assignment #8:

Out: Wednesday, April 01, 2009

Due: 2:00 pm Wednesday, April 08, 2009

In this homework you will use SRILM Toolkit¹ to generate n-gram language models and use them in the decoder. Use the setup provided at the following location: `/afs/cs.cmu.edu/project/cmt-55/lti/Courses/731/homework/HW8`. SRI LM toolkit is available in `./srilm`.

Your tasks are:

- 1) Build n-gram language models (for $n=3,4,5$) with modified Kneser-Ney discounting using English training data `./data/europarl-v3b.en`. Use the binary `./srilm/bin/i686/ngram-count`.

For additional details on usage refer to the manual on SRI LM webpage.

- 2) For each language model, calculate the Perplexity for the development set `./data/dev.en`. You can use the binary `./srilm/bin/i686/ngram` for this task.
- 3) Run MER training with each language model separately using the same setup used for HW7. Report the resulting BLEU scores. State your observations.
- 4) Run an ablation study by repeatedly splitting the corpus to generate corpus sizes of $1/2$, $1/4$, $1/8$, and $1/16$ of the full corpus. You can do this by each time selecting only the even-numbered sentences. Train for each corpus size a 3-gram LM as in Task 1 and calculate the perplexity on the development set as in Task 2. Report the perplexity values for all corpus sizes (including the full corpus) and state your observations.

Bonus Task (2 Points):

- 5) Run MER training with each of the 3-gram language models you generated in Task 4 and report the effect on BLEU score.

¹ <http://www.speech.sri.com/projects/srilm/>