100G Japanese LM

Testing result:

|  |  |  |  |
| --- | --- | --- | --- |
| ppl/wer | ECO135 | ETT | SPC160 |
| 80000 100G(new) | 13.2243/21.06 | 13.6333/26.77 | 13.2243/16.64 |
| 80000 3G(new) | 13.6551/21.34 | 14.0731/27.16 | 13.6551/17.4 |
| 80000 2G(old) | 26.0034/20.99 | 26.6318/26.54 | 26.0034/22.4 |
| 60000 2G(old) | 26.9351/21.93 | 27.5462/28.35 | 26.9351/23.34 |

This is a release of Japanese language model. I clean 50 2G Japanese corpora, and then segment Japanese word. And I generate 50 count files with them, and generate lm files with count files. Then I prune the 50 lm files to 5gram 1e-9, and interpolate them to get an overall LM. And then prune it.

Beam = 15.0

Max-active = 7000

All the models above are 3gram 1e-7.

The baseline model uses 2G(old) corpus (/work4/zhaomy/work/train\_csj/20161215/text/ALL\_mecab.txt).

The new corpus is 300G corpus.(/work5/zhaomy/corpus/Japanese)

100G out of 300G Japanese corpus is used to train the model.

The vocabulary is /work4/zhaomy/work/train\_csj/lm/vocab\_lex/20161223/vocab.

The word segmentation script is at /work4/zhaomy/work/train\_11000h/lm/20160929.

The lexicon is /work4/zhaomy/work/train\_csj/lm/vocab\_lex/20161223/lexicon.pd.

The AM is /work4/zhaomy/work/train\_csj/exp/nnet3/knowledge\_transfer/nnet\_tdnn\_no-ivector.

The test datasets are /work4/zhaomy/work/train\_csj/test\_ECO135, /work4/zhaomy/work/train\_csj/test\_ETT and /work4/zhaomy/work/train\_csj/test\_SPC160.

Conclusion:

Extension of vocabulary can improve the performance. The performance of new corpus is worse than that of the old corpus with ECO135 and ETT, but new corpus is better than old corpus with SPC160. Extension of corpus can also improve performance. But it is just a tiny improvement.