Scripts:

combine.py

Function:

* Compile all train files or test files together

Results:

* heldout\_plain.txt
* train\_plain.txt

chinese\_sw.py

English\_sw.py

Function:

* generate separate English and Chinese document for L1 and L2 according to the paper
* generate switch pairs

Results:

* english\_train.txt
* chinese\_train.txt
* chinese\_heldout.txt
* englisth\_heldout.txt
* switch (train)
* switch1 (test)
* switch\_chn (train)
* switch1\_chn (test)

ngram-count -text english\_train.txt -lm eng.lm -kndiscount3 -order 2

ngram-count -text chinese\_train.txt -lm chn.lm -kndiscount3 -order 2

Function:

* generate english or Chinese model

Results:

* eng.lm
* chn.lm

clean\_lm.py

Function:

* generate cleaned bi-grams without code-switching

Results:

* cleaned\_eng.lm
* cleaned\_chn.lm

preprocess.ipynb

Function:

* generate bigram with code-switching from English language model eng.lm and Chinese language model chn.lm

Results:

* eng\_sw.txt
* sw\_eng.txt
* chn\_sw.txt
* sw\_chn.txt

eng\_chn.py

chn\_eng.py

Function:

* generate bigrams with code-switching to replace ‘<sw>’

Results：

* eng\_chn.txt
* chn\_eng.txt

Compile all the files eng\_chn.txt, chn\_eng.txt, cleaned\_eng.lm and cleaned\_chn.lm together to generate the dual language model dlm.lm:

Middle results:

* bi\_eng.lm
* bi\_chn.lm
* uni\_eng.lm
* uni\_chn.lm

Final result:

* dlm.lm