

Build Moses on Ubuntu (64-bit) in VirtualBox and 186-server: recorded by Aaron

[<http://www.linkedin.com/in/aaronhan>]

This document introduces the record by Aaron when running Moses translation systems (<http://www.statmt.org/moses/>). If you want to get more information about the Moses please see the official website of Moses or Moses-manual.pdf (<http://www.statmt.org/moses/manual/manual.pdf>).

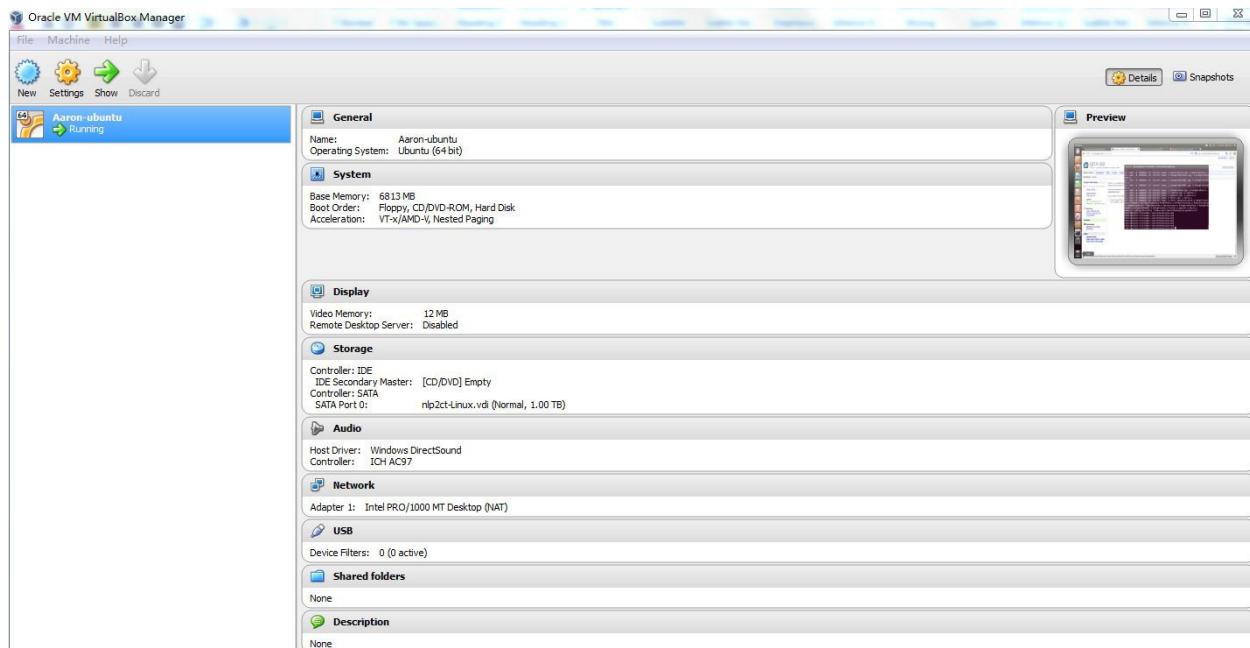
Contents

Install Moses MT system and run an example:.....	3
Another way to install GIZA++:.....	9
Another way to install IRSTLM:.....	10
Corpus preparation:.....	10
Language model training (built on the target language to ensure fluent output translation):.....	12
Training translation system (run word alignment, phrase extraction and scoring, create lexicalized reordering tables and create Moses configuration file):.....	15
Tuning translation system:.....	20
Testing the translation model:.....	25
Then we translate the testing corpus and score the translation quality using BLEU metric:.....	31
Trial of Tree to tree translation using Moses:.....	33
Build ZH->EN Tree to tree translation using Moses:.....	41
Prepare corpus:.....	41
Training language model:.....	43
Training translation model:.....	43
Testing:.....	46
Preparing larger corpus-2.....	48
Training translation model-2:.....	49
Testing-2.....	51
Calculate the testing score using BLEU:.....	53
Preparing larger corpus-2-Uni-phrase-tags.....	54
Training translation model-2-Uni-phrase-tags.....	55
Testing-2-Uni-phrase-tags.....	56
Calculate the testing score using BLEU:.....	58

Preparing larger corpus-3.....	58
Training translation model-3.....	59
Testing-3:.....	62
Calculate the testing score using BLEU:.....	65
Tuning using 100001 st -110000th sentences-3.....	65
Tuning on nlp2ct-186-sever.....	68
Preparing larger corpus-3-uni-phrse-tags.....	72
Training translation model-3-uni-phrase-tags.....	74
testing-3-uni-phrase-tags:.....	75
Calculate the testing score using BLEU:.....	77
Preparing larger corpus-4-186sever.....	77
Training translation model-4-186sever.....	77
Testing-4-186sever:.....	79
Preparing larger corpus-4-Uni-186sever.....	85
Training translation model-4-Uni-186sever.....	85
Testing-4-uni-186sever:.....	88
Build PT->CN tree-to-tree translation model with ori-phrase tags on 186-server:.....	95
Train language model:.....	95
Corpus for Training translation model:.....	101
Prepare PT corpus:.....	101
Prepare the CN corpus:.....	105
Training translation model use ori-tags:.....	106
Test translation model sue ori-tags:.....	106
Training translation model use uni-tags:.....	106
Test translation model sue uni-tags:.....	106
Tune the MT system using LEPOR metric (external metric):.....	106
Reference:.....	107

Install Moses MT system and run an example:

1. Download virtual box from <https://www.virtualbox.org/wiki/Downloads>
2. Install virtual box by double click the “VirtualBox-4.2.18-88781-Win.exe” file. [To make the Ubuntu fluent, take the following actions.
A1.download and install the VirtualBox Extension Pack outside the VM:
<https://www.virtualbox.org/wiki/Downloads>
A2.Ubuntu->Devices -> Install guest additions...]
3. Press “New” button to guide your “driver” (I use nlp2ct-Linux.vdi) into the virtual box. During the selection, chose the Ubuntu64-bit. Setting the base memory around 7GB.



4. Press the “start” to launch the Ubuntu system in the virtual box.
5. Download Giza++ (word alignment model) from <https://code.google.com/p/giza-pp/downloads/list>
6. Download IRSTLM (language model) from <http://hlt.fbk.eu/en/irstlm>
7. Download Moses (Statistical Machine Translation decoder) from <http://www.statmt.org/moses/?n=Moses.Releases>
8. Install the Giza++. If you store the Giza++ in the address Home/Aaron/Moses/giza-pp, open the “Terminal”, get into this address (press “ls” to show the items under your current position; press “cd xx” to enter the xx file; press “cd ..” to jump out current file), then press “make” command. If show “g++ command not found”, type “g++”, then type “sudo apt-get install g++” to install g++. After the install of g++, type “make” command to install Giza++ again.

```

g++ -Wall -W -DNDEBUG -O3 -funroll-loops -c KategProblemWBC.cpp -o KategProblemWBC.o
g++ -Wall -W -DNDEBUG -O3 -funroll-loops -c KategProblem.cpp -o KategProblem.o
g++ -Wall -W -DNDEBUG -O3 -funroll-loops -c StatVar.cpp -o StatVar.o
g++ -Wall -W -DNDEBUG -O3 -funroll-loops -c general.cpp -o general.o
g++ -Wall -W -DNDEBUG -O3 -funroll-loops -c mkcls.cpp -o mkcls.o
g++ -Wall -W -DNDEBUG -O3 -funroll-loops -o mkcls GDOptimization.o HCOptimization.o Problem.o IterOptimization.o ProblemTest.o RRTOptimization.o MYOptimization.o SAOptimization.o TAOptimization.o Optimization.o KategProblemTest.o KategProblemKBC.o KategProblemWBC.o KategProblem.o StatVar.o general.o mkcls.o
make[1]: Leaving directory `/home/nlp2ct/Aaron/Moses/giza-pp/mkcls-v2'

```

9. Install IRSTLM. Get into the location of the IRSTLM file. (install guideline http://sourceforge.net/apps/mediawiki/irstlm/index.php?title=Installation_Guidelines) Type “sh regenerate-makefiles.sh”. if shows “command not found aclocal failed” type “sudo apt-get install automake”, if show “libtoolize: no such file” type “sudo apt-get install libtool”. Then type “bash regenerate-makefiles.sh” again.
Type “./configure –prefix=/Home/Aaron/Moses/irstlm-5.80.03” to generate and locate the “Makefile”

```

checking getopt.h usability... yes
checking getopt.h presence... yes
checking for getopt.h... yes
configure: creating ./config.status
config.status: creating Makefile
config.status: creating src/Makefile
config.status: creating scripts/Makefile
config.status: creating config.h
config.status: executing depfiles commands
config.status: executing libtool commands
configure: The software will be installed into /Home/Aaron/Moses/irstlm-5.80.03
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/irstlm-5.80.03$ █

```

Type “make” for compilation. If show “zlib.h: No such file” type “sudo apt-get install zlib1g-dev” to install zlib. Type “make” again.

```

make[2]: Leaving directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03/src'
Making all in scripts
make[2]: Entering directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03/scripts'
make[2]: Nothing to be done for `all'.
make[2]: Leaving directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03/scripts'
make[2]: Entering directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03'
make[2]: Leaving directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03'
make[1]: Leaving directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03'
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/irstlm-5.80.03$ █

```

Type “sudo make install” for installation.

```
make[2]: Leaving directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03/scripts'
make[1]: Leaving directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03/scripts'
make[1]: Entering directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03'
make[2]: Entering directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03'
make[2]: Nothing to be done for `install-exec-am'.
make[2]: Nothing to be done for `install-data-am'.
make[2]: Leaving directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03'
make[1]: Leaving directory `/home/nlp2ct/Aaron/Moses/irstlm-5.80.03'
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/irstlm-5.80.03$ █
```

The IRSTLM library and commands are generated respectively under the address
“/Home/Aaron/Moses/irstlm-5.80.03”

10. Install Boost (C++ libraries).

Download it from (http://www.boost.org/users/history/version_1_52_0.html).

```
cd boost_1_52_0
./bootstrap.sh --prefix=/Home/Aaron/Moses/boost_1_52_0
```

```
Building Boost.Build engine with toolset gcc... tools/build/v2/engine/bin.linuxx
86_64/b2
Detecting Python version... 2.7
Detecting Python root... /usr
Unicode/ICU support for Boost.Regex?... /usr
Generating Boost.Build configuration in project-config.jam...

Bootstrapping is done. To build, run:
```

```
./b2
```

```
To adjust configuration, edit 'project-config.jam'.
Further information:
```

- Command line help:

```
./b2 --help
```
- Getting started guide:
http://www.boost.org/more/getting_started/unix-variants.html
- Boost.Build documentation:
<http://www.boost.org/boost-build2/doc/html/index.html>

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/boost_1_52_0$ █
```

Type “./b2” to install boost.

```
gcc.compile.c++ bin.v2/libs/wave/build/gcc-4.6/release/link-static/threading-multi/wave
_config_constant.o
common.mkdir bin.v2/libs/wave/build/gcc-4.6/release/link-static/threading-multi/cpplexe
r
common.mkdir bin.v2/libs/wave/build/gcc-4.6/release/link-static/threading-multi/cpplexe
r/re2clex
gcc.compile.c++ bin.v2/libs/wave/build/gcc-4.6/release/link-static/threading-multi/cppl
exer/re2clex/aq.o
gcc.compile.c++ bin.v2/libs/wave/build/gcc-4.6/release/link-static/threading-multi/cppl
exer/re2clex/cpp_re.o
gcc.archive bin.v2/libs/wave/build/gcc-4.6/release/link-static/threading-multi/libboost
_wave.a
common.copy stage/lib/libboost_wave.a
...failed updating 2 targets...
...skipped 6 targets...
...updated 971 targets...
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/boost_1_52_0$
```

11. Install other dependencies (gcc, zlib, bzip2).

Type “sudo apt-get install build-essential libbz-dev libbz2-dev” to install.

```
Setting up libtimedate-perl (1.2000-1) ...
Setting up libdpkg-perl (1.16.1.2ubuntu7.1) ...
Setting up dpkg-dev (1.16.1.2ubuntu7.1) ...
Setting up build-essential (11.5ubuntu2.1) ...
Setting up fakeroot (1.18.2-1) ...
update-alternatives: using /usr/bin/fakeroot-sysv to provide /usr/bin/fakeroot (fakeroo
t) in auto mode.
Setting up libalgorithm-diff-perl (1.19.02-2) ...
Setting up libalgorithm-diff-xs-perl (0.04-2build2) ...
Setting up libalgorithm-merge-perl (0.08-2) ...
Setting up libbz2-dev (1.0.6-1) ...
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses$
```

12. Install Moses.

Type “git clone git://github.com/moses-smt/mosesdecoder.git”.

```
Cloning into 'mosesdecoder'...
remote: Counting objects: 90550, done.
remote: Compressing objects: 100% (23416/23416), done.
Receiving objects: 100% (90550/90550), 111.65 MiB | 1.01 MiB/s, done.
remote: Total 90550 (delta 68741), reused 86391 (delta 64660)
Resolving deltas: 100% (68741/68741), done.
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses$
```

Compile Moses. To examine the options you want, type “cd ~/mosesdecoder”, “./bjam --help”.

There will be automatic updates.

```
*** No errors detected
gcc.compile.c++ mert/bin/gcc-4.6/release/debug-symbols-on/link-static/threading-multi/VocabularyTest.o
gcc.link mert/bin/gcc-4.6/release/debug-symbols-on/link-static/threading-multi/vocabulary_test
testing.unit-test mert/bin/gcc-4.6/release/debug-symbols-on/link-static/threading-multi/vocabulary_test.passed
Running 2 test cases...

*** No errors detected
common.copy mert/mert
common.copy mert/extractor
common.copy mert/evaluator
common.copy mert/pro
common.copy mert/kbmira
common.copy mert/sentence-bleu
...updated 746 targets...
SUCCESS
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/mosesdecoder$ █
```

13. Run Moses with an example.

Type the following commands to run the example:

```
cd ~/mosesdecoder
wget http://www.statmt.org/moses/download/sample-models.tgz
tar xzvf sample-models.tgz
cd sample-models
~/Aaron/Moses/mosesdecoder/bin/moses -f phrase-model/moses.ini < phrase-model/in > out
```

```
Defined parameters (per moses.ini or switch):
config: phrase-model/moses.ini
feature: KENLM name=LM factor=0 order=3 num-features=1 path=lm/europarl.srilm.gz Distortion WordPenalty UnknownWordPenalty PhraseDictionaryMemory input-factor=0 output-factor=0 path=phrase-model/phrase-table num-features=1 table-limit=10
input-factors: 0
mapping: T 0
n-best-list: nbest.txt 100
weight: WordPenalty=0 LM= 1 Distortion= 1 PhraseDictionaryMemory= 1
/home/nlp2ct/Aaron/Moses/mosesdecoder/bin
line=KENLM name=LM factor=0 order=3 num-features=1 path=lm/europarl.srilm.gz
FeatureFunction: LM start: 0 end: 0
Loading the LM will be faster if you build a binary file.
Reading lm/europarl.srilm.gz
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100
**The ARPA file is missing <unk>. Substituting log10 probability -100.000.
*****
line=Distortion
FeatureFunction: Distortion0 start: 1 end: 1
line=WordPenalty
FeatureFunction: WordPenalty0 start: 2 end: 2
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 3 end: 3
line=PhraseDictionaryMemory input-factor=0 output-factor=0 path=phrase-model/phrase-table num-features=1 table-limit=10
FeatureFunction: PhraseDictionaryMemory0 start: 4 end: 4
Start loading text SCFG phrase table. Moses format : [1.000] seconds
Reading phrase-model/phrase-table
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100
*****
IO from STDOUT/STDIN
Created input-output object : [1.000] seconds
Translating line 0 in thread id 140710001190656
Translating: das ist ein kleines haus
Line 0: Collecting options took 0.000 seconds
Line 0: Search took 0.000 seconds
BEST TRANSLATION: this is a small house [11111] [total=-28.923] core=(-27.091,0.000,-5.000,0.000,-1.833)
Line 0: Translation took 0.000 seconds total
Translating line 1 in thread id 140710001190656
Translating: das ist ein kleines haus
Line 1: Collecting options took 0.000 seconds
Line 1: Search took 0.000 seconds
BEST TRANSLATION: this is a small house [11111] [total=-28.923] core=(-27.091,0.000,-5.000,0.000,-1.833)
Line 1: Translation took 0.000 seconds total
Name:moses VmPeak:197672 kB VmRSS:32300 kB RSSMax:32544 kB user:0.648 sys:0.028 CPU:0.676 real:0.783
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/mosesdecoder/sample-models$ █
```

The translation results of source sentence “das ist ein kleines haus” will be shown in the file of out as “it is a small house”. Succeed!

To test the Chart decoder, type the following command, the translation result will be shown in out.stt file:

```
~/Aaron/Moses/mosesdecoder/bin/moses_chart -f string-to-tree/moses.ini < string-to-tree/in >  
out.stt
```

```
Reading string-to-tree/rule-table  
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80  
---85---90---95--100  
*****  
*****  
max-chart-span: 20  
IO from STDOUT/STDIN  
Created input-output object : [0.000] seconds  
Translating: <s> das ist ein kleines haus </s> ||| [0,0]=X (1) [0,1]=X (1) [0,2]  
]=X (1) [0,3]=X (1) [0,4]=X (1) [0,5]=X (1) [0,6]=X (1) [1,1]=X (1) [1,2]=X (1)  
[1,3]=X (1) [1,4]=X (1) [1,5]=X (1) [1,6]=X (1) [2,2]=X (1) [2,3]=X (1) [2,4]=X  
(1) [2,5]=X (1) [2,6]=X (1) [3,3]=X (1) [3,4]=X (1) [3,5]=X (1) [3,6]=X (1) [4,4]=X (1) [4,5]=X (1) [4,6]=X (1) [5,5]=X (1) [5,6]=X (1) [6,6]=X (1)  
  
0   1   2   3   4   5   6  
0   3   2   2   2   1   0  
0   0   0   0   0   0  
0   0   0   3   0  
0   0   4   0  
0   4   0  
0   0  
1  
BEST TRANSLATION: 41 TOP -> <s> S </s> :1-1 : c=-3.206 core=(-6.413,-2.000,0.00  
0,0.000) [0..6] 20 [total=-15.501] core=(-27.091,-7.000,0.000,-3.912)  
Translation took 0.000 seconds  
End. : [0.000] seconds  
Name:moses_chart      VmPeak:190436 kB      VmRSS:31856 kB  RSSMax:32312 kB  
ser:0.596      sys:0.048      CPU:0.644      real:0.795  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/mosesdecoder/sample-models$ █
```

To test the tree-to-tree demo, type the following command, the translation result will be shown in out.ttt file:

```
~/Aaron/Moses/mosesdecoder/bin/moses_chart -f tree-to-tree/moses.ini < tree-to-tree/in.xml  
> out.ttt
```

```

[3,8]=X (1) [3,9]=X (1) [3,10]=X (1) [3,11]=X (1) [3,12]=X (1) [3,13]=X (1) [4,4]=X (1)
[4,5]=X (1) [4,6]=X (1) [4,7]=X (1) [4,8]=X (1) [4,9]=X (1) [4,10]=X (1) [4,11]=X (1)
[4,12]=X (1) [4,13]=X (1) [5,5]=X (1) [5,6]=X (1) [5,7]=X (1) [5,8]=X (1) [5,9]=X (1)
[5,10]=X (1) [5,11]=X (1) [5,12]=X (1) [5,13]=X (1) [6,6]=X (1) [6,7]=X (1) [6,7]=N
S (1) [6,8]=X (1) [6,9]=X (1) [6,10]=X (1) [6,11]=X (1) [6,12]=X (1) [6,13]=X (1) [7,7]
]=X (1) [7,7]=NS (1) [7,8]=X (1) [7,9]=X (1) [7,10]=X (1) [7,11]=X (1) [7,12]=X (1) [7
,13]=X (1) [8,8]=X (1) [8,9]=X (1) [8,10]=X (1) [8,11]=X (1) [8,12]=X (1) [8,12]=OS (1
) [8,13]=X (1) [9,9]=X (1) [9,10]=X (1) [9,11]=X (1) [9,12]=X (1) [9,13]=X (1) [10,10]
=X (1) [10,11]=X (1) [10,12]=X (1) [10,13]=X (1) [11,11]=X (1) [11,12]=X (1) [11,13]=X
(1) [12,12]=X (1) [12,13]=X (1) [13,13]=X (1)

      0   1   2   3   4   5   6   7   8   9   10  11  12  13
 1  1   1   1   1   1   1   1   1   1   1   1   1   1   0
   1   0   1   0   0   0   1   0   0   0   0   0   0   0   0
   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
   1   0   0   1   0   0   0   0   0   0   0   1   0
   1   0   1   0   0   0   0   0   0   0   0   0
   1   0   0   0   0   0   0   0   0   0   0
   1   0   0   0   0   0   0   0   0   0
   1   0   0   0   0   0   0   0   0
   1   0   0   0   0   0   0   0
   1   0   0   0   0
   1   0   0
   1   0

BEST TRANSLATION: 35 S -> S </s> :0-0 : c=0.000 core=(-1.000,0.000,0.000,0.000,0.000,
0.000,0.000,0.000) [0..13] 33 [total=-9.680] core=(-9.000,0.000,-5.839,-9.522,0.000,
6.000,4.999,3.000)
Translation took 0.000 seconds
End. : [0.000] seconds
Name:moses_chart          VmPeak:160972 kB          VmRSS:2720 kB    RSSMax:3520 kB  user:0
.008    sys:0.008        CPU:0.016        real:0.075
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/mosesdecoder/sample-models$ █

```

Another way to install GIZA++:

Type the following commands to install GIZA++.

```
wget http://giza-pp.googlecode.com/files/giza-pp-v1.0.7.tar.gz
```

```
tar xzvf giza-pp-v1.0.7.tar.gz
```

```
cd giza-pp
```

make

above commands will create the binaries `~/giza-pp/GIZA++-v2/GIZA++`, `~/giza-pp/GIZA++-v2/snt2cooc.out` and `~/giza-pp/mkcls-v2/mkcls`. To automatically copy these files to somewhere that Moses can find, type the following commands:

```
cd ~/mosesdecoder  
mkdir tools
```

```
cp ~/Aaron/Moses/giza-pp/GIZA++-v2/GIZA++ ~/Aaron/Moses/giza-pp/GIZA++-v2/snt2cooc.out  
~/Aaron/Moses/giza-pp/mkcls-v2/mkcls tools
```

above commands will copy the three binaries “GIZA++, snt2cooc.out, and mkcls” into the directory “~/mosesdecoder/tools”.

When you come to run the training, you need to tell the training script where GIZA++ was installed using the –external-bin-dir argument as below:

```
train-model.perl –external-bin-dir $HOME/mosesdecoder/tools
```

Another way to install IRSTLM:

Download the latest version of IRSTLM.

```
tar zxvfirstlm-5.80.03.tgz  
cdirstlm-5.80.03  
. /regenerate-makefiles.sh  
. /configure –prefix=$HOME/Aaron/Moses/irstlm-5.80.03  
sudo make install
```

Corpus preparation:

```
mkdir corpus  
cd corpus  
wget http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz  
tar zxvf training-parallel-nc-v8.tgz
```

corpus tokenization (add spaces between words and punctuations):

type the following command for corpus tokenization:

```
~/Aaron/Moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en <  
~/Aaron/corpus/training/news-commentary-v8.fr-en.en > ~/Aaron/corpus/news-commentary-  
v8.fr-en.tok.en
```

```
~/Aaron/Moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l fr <  
~/Aaron/corpus/training/news-commentary-v8.fr-en.fr > ~/Aaron/corpus/news-commentary-  
v8.fr-en.tok.fr
```

The tokenized files will be generated in the rectory file “~/Aaron/corpus”

Truecasing (to reduce data sparsity, convert the initial words of the sentence into the most probable casing):

To get the truecasing training models that are the statistics data extracted from text, type the following commands:

```
~/Aaron/Moses/mosesdecoder/scripts/recaser/train-truecaser.perl --model  
~/Aaron/corpus/trucase-model.en --corpus ~/Aaron/corpus/news-commentary-v8.fr-en.tok.en
```

```
~/Aaron/Moses/mosesdecoder/scripts/recaser/train-truecaser.perl --model  
~/Aaron/corpus/trucase-model.fr --corpus ~/Aaron/corpus/news-commentary-v8.fr-en.tok.fr
```

Above commands will generate the model files “truecase-model.en” and “truecase-model.fr” under the directory “/Aaron/corpus”

Using the extracted truecasing training models to perform the truecase function as below commands:

```
~/Aaron/Moses/mosesdecoder/scripts/recaser/truecase.perl --model ~/Aaron/corpus/trucase-  
model.en < ~/Aaron/corpus/news-commentary-v8.fr-en.tok.en > ~/Aaron/corpus/news-  
commentary-v8.fr-en.true.en
```

```
~/Aaron/Moses/mosesdecoder/scripts/recaser/truecase.perl --model ~/Aaron/corpus/trucase-  
model.fr < ~/Aaron/corpus/news-commentary-v8.fr-en.tok.fr > ~/Aaron/corpus/news-  
commentary-v8.fr-en.true.fr
```

Above commands will generate the files “news-commentary-v8.fr-en.true.ed” and “news-commentary-v8.fr-en.true.fr” under directory “Aaron/corpus”

Cleaning (to remove the mis-aligned sentences, long sentences and empty sentences, which may cause problems with the training pipeline)

Type the following command to delete the sentences whose length is larger than 80:

```
~/Aaron/Moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/Aaron/corpus/news-  
commentary-v8.fr-en.true fr en ~/Aaron/corpus/news-commentary-v8.fr-en.clean 1 80
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/mosesdecoder$ ~/Aaron/Moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/Aaron/corpus/news-commentary-v8.fr-en.true fr en ~/Aaron/corpus/news-commentary-v8.fr-en.clean 1 80  
clean-corpus.perl: processing /home/nlp2ct/Aaron/corpus/news-commentary-v8.fr-en.true.fr & .en to /home/nlp2ct/Aaron/corpus/news-commentary-v8.fr-en.clean, cutoff 1-80  
.....(100000).....  
Input sentences: 157168 Output sentences: 155362  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/mosesdecoder$ █
```

The files “news-commentary-v8.fr-en.clean.en” and “news-commentary-v8.fr-en.clean.fr” will be generated in the directory “~/Aaron/corpus”.

=====

Language model training (built on the target language to ensure fluent output translation):

```
cd ~/Aaron  
mkdir lm  
cd lm  
~/Aaron/Moses/irstlm-5.80.03/scripts/add-start-end.sh < ~/Aaron/corpus/news-commentary-v8.fr-en.true.en > news-commentary-v8.fr-en.sb.en
```

The above commands will generate the file “news-commentary-v8.fr-en.sb.en” in the directory “Aaron/lm”. In this file, each sentence is added with the start and end symbol “<s> </s>”.

Type the following command to generate the language model of English.
export IRSTLM=\$Home/Aaron/Moses/irstlm-5.80.03; ~/Aaron/Moses/irstlm-5.80.03/scripts/build-lm.sh -i news-commentary-v8.fr-en.sb.en -t ./tmp -p -s improved-kneserney -o news-commentary-v8.fr-en.lm.en

```
Temporary directory ./tmp does not exist  
creating ./tmp  
Extracting dictionary from training corpus  
Splitting dictionary into 3 lists  
Extracting n-gram statistics for each word list  
Important: dictionary must be ordered according to order of appearance of words in data  
used to generate n-gram blocks, so that sub language model blocks results ordered too  
dict.000  
dict.001  
dict.002  
$bin/ngt -i="$inpfle" -n=$order -goout=y -o="$gzip -c > $tmpdir/ngram.${sdict}.gz" -  
fd="$tmpdir/$sdict" $dictionary -iknstat="$tmpdir/ikn.stat.$sdict" >> $logfile 2>&1  
Estimating language models for each word list  
dict.000  
dict.001  
dict.002  
$scr/build-sublm.pl $verbose $prune $smoothing "cat $tmpdir/ikn.stat.dict.*" --size $order --ngrams "$gunzip -c $tmpdir/ngram.${sdict}.gz" -sublm $tmpdir/lm.$sdict >> $logfile 2>&1  
Merging language models into news-commentary-v8.fr-en.lm.en  
Cleaning temporary directory ./tmp  
Removing temporary directory ./tmp  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/lm$ █
```

Above command generate the file “news-commentary-v8.fr-en.lm.en.gz” in the directory “Aaron/lm”.

Type the following command to generate the file “news-commentary-v8.fr-en.arpa.en” in the directory “Aaron/lm”

```
~/Aaron/Moses/irstlm-5.80.03/src/compile-lm -text news-commentary-v8.fr-en.lm.en.gz news-commentary-v8.fr-en.arpa.en
```

```
infile: news-commentary-v8.fr-en.lm.en.gz
outfile: news-commentary-v8.fr-en.arpa.en
loading up to the LM level 1000 (if any)
dub: 10000000
Language Model Type of news-commentary-v8.fr-en.lm.en.gz is 1
Language Model Type is 1
iARPA
loadtxt_ram()
1-grams: reading 62504 entries
done level 1
2-grams: reading 906385 entries
done level 2
3-grams: reading 378679 entries
done level 3
done
OOV code is 62503
OOV code is 62503
Saving in txt format to news-commentary-v8.fr-en.arpa.en
savetxt: news-commentary-v8.fr-en.arpa.en
save: 62504 1-grams
save: 906385 2-grams
save: 378679 3-grams
done
nlp2ct@nlp2ct-VirtualBox:~/Aaron/lm$
```

To make the faster loading, type the following command to binary the arpa.en file
“~/Aaron/Moses/mosesdecoder/bin/build_binary news-commentary-v8.fr-en.arpa.en news-commentary-v8.fr-en.blm.en”

To check the language model, type the following command: “\$ echo “is this an English sentence ?” | ~/Aaron/Moses/mosesdecoder/bin/query news-commentary-v8.fr-en.arpa.en”. it will show the following result.

```
Loading the LM will be faster if you build a binary file.
Reading news-commentary-v8.fr-en.arpa.en
-----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---
-90---95--100
*****$: command not found
*****
Loading statistics:
Name:query      VmPeak:47816 kB VmRSS:31108 kB   RSSMax:31940 kB user:0.452028    sys:0.0
12      CPU:0.464028    real:0.597656
After queries:
Name:query      VmPeak:47816 kB VmRSS:31108 kB   RSSMax:31940 kB user:0.452028    sys:0.0
12      CPU:0.464028    real:0.601562
Total time including destruction:
Name:query      VmPeak:47816 kB VmRSS:1724 kB   RSSMax:31940 kB user:0.452028    sys:0.0
12      CPU:0.464028    real:0.601562
nlp2ct@nlp2ct-VirtualBox:~/Aaron/lm$
```

If using the following command: “\$ echo “is this an English sentence ?” | ~/Aaron/Moses/mosesdecoder/bin/query news-commentary-v8.fr-en.blm.en”. it will show the following result:

```
>Loading statistics:  
Name:query      VmPeak:46784 kB VmRSS:30828 kB RSSMax:30828 kB user:0 sys:0.004 CPU:0.004 real:0.0078125  
$: command not found  
After queries:  
Name:query      VmPeak:46788 kB VmRSS:30828 kB RSSMax:30828 kB user:0 sys:0.004 CPU:0.004 real:0.0820312  
Total time including destruction:  
Name:query      VmPeak:46788 kB VmRSS:1516 kB RSSMax:30828 kB user:0.004 sys:0.004 CPU:0.008 real:0.0859375  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/lm$
```

=====

Training translation system (run word alignment, phrase extraction and scoring, create lexicalized reordering tables and create Moses configuration file):

Type the following command to train the translation system using language model:

```
cd ~/Aaron  
mkdir working  
cd working  
~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train -corpus  
~/Aaron/corpus/news-commentary-v8.fr-en.clean -f fr -e en -alignment grow-diag-final-and  
-reordering msd-bidirectional-fe -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8  
-external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out &
```

```
~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train -corpus ~/Aaron/corpus/news-commentary-v8.fr-en.clean -f fr -e en -alignment grow-diag-final-and -reordering msd-bidirectional-fe -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out &
```

The running details are written into the ~/working/training.out file step by step as below:

```

Using SCRIPTS_ROOTDIR: /home/nlp2ct/Aaron/Moses/mosesdecoder/scripts
Using single-thread GIZA
(1) preparing corpus @ Fri Oct 25 19:30:19 CST 2013|
Executing: mkdir -p /home/nlp2ct/Aaron/working/train/corpus
(1.0) selecting factors @ Fri Oct 25 19:30:19 CST 2013
(1.1) running mkcls @ Fri Oct 25 19:30:19 CST 2013
/home/nlp2ct/Aaron/Moses/mosesdecoder/tools/mkcls -c50 -n2 -p/home/nlp2ct/Aaron/corpus/news-commentary-
v8.fr-en.clean.fr -V/home/nlp2ct/Aaron/working/train/corpus/fr.vcb.classes opt
Executing: /home/nlp2ct/Aaron/Moses/mosesdecoder/tools/mkcls -c50 -n2 -p/home/nlp2ct/Aaron/corpus/news-
commentary-v8.fr-en.clean.fr -V/home/nlp2ct/Aaron/working/train/corpus/fr.vcb.classes opt

***** 2 runs. (algorithm:TA)*****
;KategProblem:cats: 50 words: 71168

start-costs: MEAN: 7.50606e+07 (7.50496e+07-7.50716e+07) SIGMA:10981
end-costs: MEAN: 6.96206e+07 (6.9613e+07-6.96283e+07) SIGMA:7614.14
start-pp: MEAN: 656.723 (655.266-658.18) SIGMA:1.45712
end-pp: MEAN: 218.785 (218.449-219.122) SIGMA:0.336596
iterations: MEAN: 1.92174e+06 (1.88842e+06-1.95504e+06) SIGMA:33310
time: MEAN: 117.479 (114.747-120.212) SIGMA:2.73217
(1.1) running mkcls @ Fri Oct 25 19:34:46 CST 2013
/home/nlp2ct/Aaron/Moses/mosesdecoder/tools/mkcls -c50 -n2 -p/home/nlp2ct/Aaron/corpus/news-commentary-
v8.fr-en.clean.en -V/home/nlp2ct/Aaron/working/train/corpus/en.vcb.classes opt

```

If you want to see the running thread, type command “top” to show the running details in the window as below:

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
2610	nlp2ct	20	0	381m	368m	2028	R	94.1	5.6	28:22.37	GIZA++
1893	nlp2ct	20	0	905m	280m	41m	S	2.0	4.2	1:26.69	firefox
1969	nlp2ct	20	0	398m	34m	22m	S	1.7	0.5	0:47.25	plugin-containe
1163	root	20	0	374m	258m	22m	S	0.7	3.9	0:40.56	Xorg
1536	nlp2ct	20	0	109m	1572	1016	S	0.7	0.0	0:14.23	VBoxClient
2128	nlp2ct	20	0	519m	18m	11m	S	0.3	0.3	0:06.06	gnome-terminal
1	root	20	0	24460	2356	1368	S	0.0	0.0	0:00.96	init
2	root	20	0	0	0	0	S	0.0	0.0	0:00.00	kthreadd
3	root	20	0	0	0	0	S	0.0	0.0	0:00.12	ksoftirqd/0

Type command “ctrl+c” to shut down the current running (top) shown in the window as below [type “ctrl+z” to pause the run; type “jobs” show the running; type “bg + job-num” to put the detail background; type “fg + job-num” to show running thread (bring to fore-ground)]:

17	root	20	0	0	0	0	S	0.0	0.0	0:00.55	khubdd
18	root	0	-20	0	0	0	S	0.0	0.0	0:00.00	md
21	root	20	0	0	0	0	S	0.0	0.0	0:00.00	khungtaskd
22	root	20	0	0	0	0	S	0.0	0.0	0:00.00	kswapd0
23	root	25	5	0	0	0	S	0.0	0.0	0:00.00	ksmd
24	root	39	19	0	0	0	S	0.0	0.0	0:00.00	khugepaged

Around 2 hours and 10 minutes later, the translation training stage will be finished. By command “ctrl+c”, the following figure show the cmd window content:

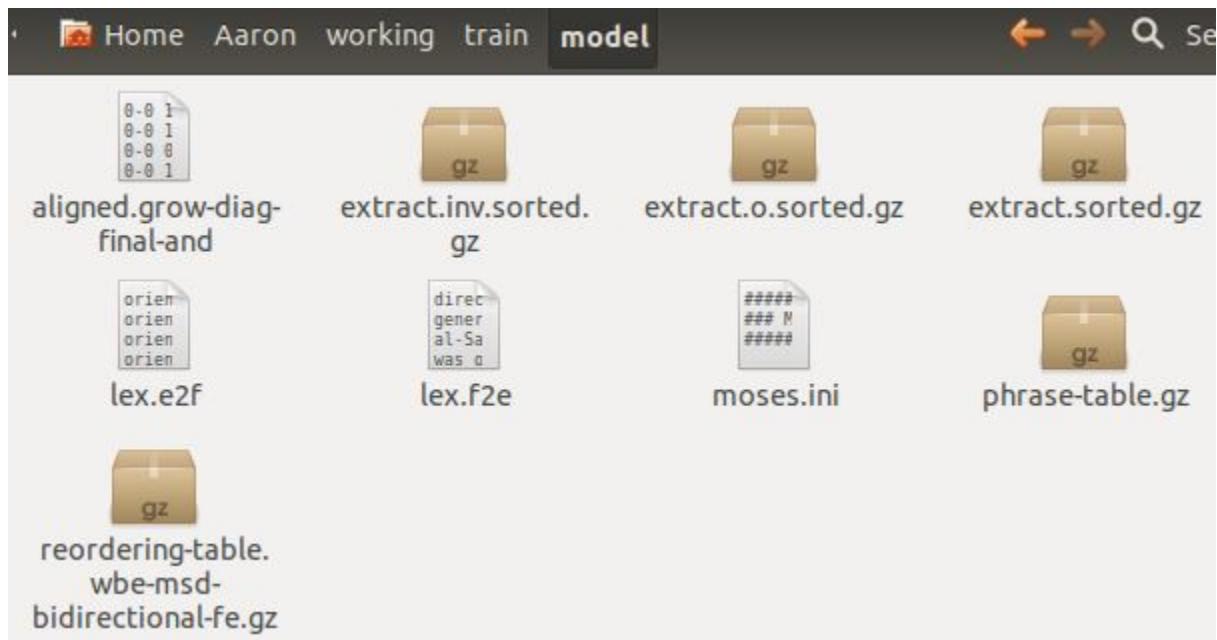
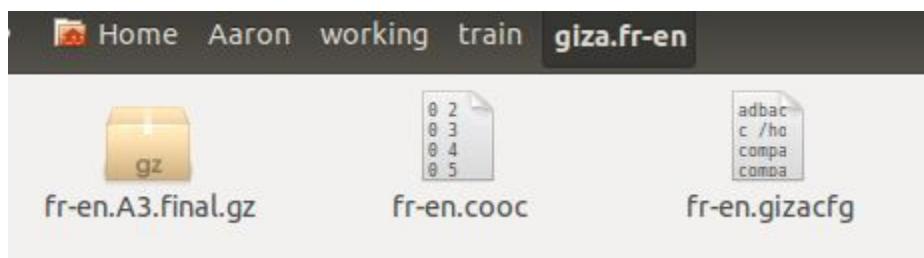
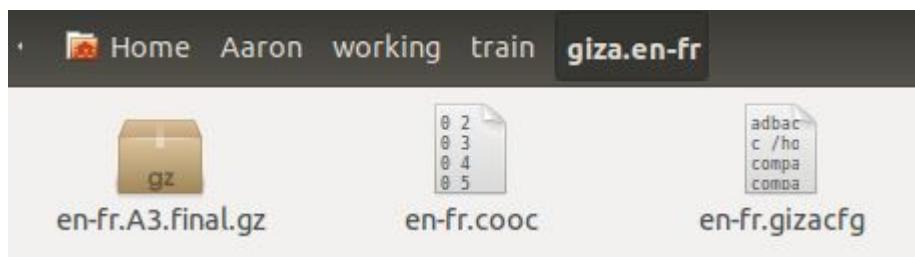
PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
1163	root	20	0	382m	266m	22m	S	1.0	4.0	1:01.29	Xorg
1893	nlp2ct	20	0	918m	263m	35m	S	1.0	4.0	3:13.69	firefox
1969	nlp2ct	20	0	398m	36m	22m	S	0.7	0.5	2:24.60	plugin-containe
2128	nlp2ct	20	0	519m	17m	10m	S	0.7	0.3	0:12.28	gnome-terminal
1536	nlp2ct	20	0	109m	1564	1008	S	0.3	0.0	0:43.70	VBoxClient
1543	nlp2ct	20	0	27924	3644	620	S	0.3	0.1	0:03.17	dbus-daemon
1607	nlp2ct	20	0	1011m	34m	17m	S	0.3	0.5	0:17.61	nautilus
1613	nlp2ct	20	0	415m	11m	7768	S	0.3	0.2	0:03.01	bamfdaemon
2679	nlp2ct	20	0	17340	1312	948	R	0.3	0.0	0:07.43	top
1	root	20	0	24460	1464	1368	S	0.0	0.0	0:00.96	init
2	root	20	0	0	0	0	S	0.0	0.0	0:00.00	kthreadd
3	root	20	0	0	0	0	S	0.0	0.0	0:00.26	ksoftirqd/0
5	root	20	0	0	0	0	S	0.0	0.0	0:00.32	kworker/u:0
6	root	RT	0	0	0	0	S	0.0	0.0	0:00.00	migration/0
7	root	RT	0	0	0	0	S	0.0	0.0	0:00.16	watchdog/0
8	root	0	-20	0	0	0	S	0.0	0.0	0:00.00	cpuset
9	root	0	-20	0	0	0	S	0.0	0.0	0:00.00	khelper
10	root	20	0	0	0	0	S	0.0	0.0	0:00.00	kdevtmpfs
11	root	0	-20	0	0	0	S	0.0	0.0	0:00.00	netns
12	root	20	0	0	0	0	S	0.0	0.0	0:00.02	sync_supers
13	root	20	0	0	0	0	S	0.0	0.0	0:00.00	bdi-default
14	root	0	-20	0	0	0	S	0.0	0.0	0:00.00	kintegrityd
15	root	0	-20	0	0	0	S	0.0	0.0	0:00.00	kblockd
16	root	0	-20	0	0	0	S	0.0	0.0	0:00.00	ata_sff
17	root	20	0	0	0	0	S	0.0	0.0	0:00.35	khubd
18	root	0	-20	0	0	0	S	0.0	0.0	0:00.00	md
21	root	20	0	0	0	0	S	0.0	0.0	0:00.00	khungtaskd
22	root	20	0	0	0	0	S	0.0	0.0	0:00.42	kswapd0
23	root	25	5	0	0	0	S	0.0	0.0	0:00.00	ksmd
24	root	39	19	0	0	0	S	0.0	0.0	0:00.00	khugepaged
[2]+	Done										~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train -corpus ~/Aaron/corpus/news-commentary-v8.fr-en.clean -f fr -e en -alignment grow-diag-final-and -reordering msd-bidirectional-fe -lm 0:3:\$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools &>training.out
											nlp2ct@nlp2ct-VirtualBox:~/Aaron/working\$ █

The following content is shown in the end of the file “~/working/training.out”:

Type “fg 2”, it will show that the progress is finished:

```
 24 root      39 19      0      0      0 S  0.0  0.0   0:00.00 khugepaged
[2]+ Done                  ~Aaron/Moses/mosesdecoder/scripts/training/train-
model.perl -root-dir train -corpus ~Aaron/corpus/news-commentary-v8.fr-en.clean
-f fr -e en -alignment grow-diag-final-and -reordering msd-bidirectional-fe -lm
0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~Aaron/
Moses/mosesdecoder/tools &>training.out
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ fg 2
bash: fg: 2: no such job
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ █
```

There will be a “train” file generated at “~/Aaron/working/train”. The “train” file contains 4 files including “corpus, giza.en-fr, giza.fr-en, and model”. The 4 files contain the below files respectively:



```
=====
```

Tuning translation system:

The weights used by Moses to weight the different models against each other are not optimized, as shown in the moses.ini file. To tune the parameters, it requires a small amount of parallel data, which is separated from the training data.

First, we download the WMT08 data, tokenise and truecase the corpus. The WMT08 corpus is used as development set in WMT12.

Make a new file “tune” to store the tuning corpus, then

```
cd ~/corpus/tune  
wget http://www.statmt.org/wmt12/dev.tgz  
tar zxvf dev.tgz  
  
~/Aaron/corpus/tune$ ~/Aaron/Moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en <  
dev/news-test2008.en > news-test2008.tok.en  
  
~/Aaron/corpus/tune$ ~/Aaron/Moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l fr <  
dev/news-test2008.en > news-test2008.tok.fr  
  
~/Aaron/Moses/mosesdecoder/scripts/recaser/truecase.perl --model ~/Aaron/corpus/truecase-  
model.en < news-test2008.tok.en > news-test2008.true.en  
  
~/Aaron/Moses/mosesdecoder/scripts/recaser/truecase.perl --model ~/Aaron/corpus/truecase-  
model.fr < news-test2008.tok.fr > news-test2008.true.fr
```

After above commands, there will be the following files generated for tuning:



Now, we begin the tuning stage using the MERT (minimum error rate training) method.

```
cd ~/Aaron/working  
~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/news-test2008.true.fr ~/Aaron/corpus/tune/news-test2008.true.en  
~/Aaron/Moses/mosesdecoder/bin/moses train/model/moses.ini --mertdir  
~/Aaron/Moses/mosesdecoder/bin/ &> mert.out &
```

```
top,  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/news-test2008.true.fr ~/Aaron/corpus/tune/news-test2008.true.en ~/Aaron/Moses/mosesdecoder/bin/moses train/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ &> mert.out &  
[1] 12434  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ top  
  
top - 12:26:23 up 2:06, 2 users, load average: 1.14, 0.72, 0.33  
Tasks: 153 total, 1 running, 152 sleeping, 0 stopped, 0 zombie  
Cpu(s): 98.3%us, 1.3%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.3%si, 0.0%st  
Mem: 6786756k total, 4410596k used, 2376160k free, 214852k buffers  
Swap: 7336956k total, 0k used, 7336956k free, 2338496k cached  
  
 PID USER      PR  NI    VIRT    RES    SHR S %CPU %MEM     TIME+   COMMAND  
12470 nlp2ct    20    0  678m 595m  31m S 93.1  9.0  1:51.71 moses  
23689 nlp2ct    20    0 1254m 401m  53m S  3.0  6.1  2:02.51 firefox
```

Type “Ctrl+c” to exit the showing, “jobs” to show the job number, “ctrl+z” to pause the job, “fg [num_job]” to show the running job foreground, “ps” to show running, the jobs are paused as below:

```
8 root      0 -20      0      0 S  0.0  0.0  0:00.00 cpuset  
9 root      0 -20      0      0 S  0.0  0.0  0:00.00 khelper  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ jobs  
[1]+  Running                  ~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/news-test2008.true.fr ~/Aaron/corpus/tune/news-test2008.true.en ~/Aaron/Moses/mosesdecoder/bin/moses train/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ &>mert.out &  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ fg 1  
~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/news-test2008.true.fr ~/Aaron/corpus/tune/news-test2008.true.en ~/Aaron/Moses/mosesdecoder/bin/moses train/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ &>mert.out  
C^Cnlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ jobs  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ps  
  PID TTY      TIME CMD  
12162 pts/3    00:00:00 bash  
12504 pts/3    00:00:00 ps
```

To save the time, type the following command to run the tuning with 6 threads,

```
~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/news-test2008.true.fr ~/Aaron/corpus/tune/news-test2008.true.en
~/Aaron/Moses/mosesdecoder/bin/moses train/model/moses.ini --mertdir
~/Aaron/Moses/mosesdecoder/bin/ --decoder-flags="-threads 6" &> mert.out &
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/news-test2008.true.fr ~/Aaron/corpus/tune/news-test2008.true.en ~/Aaron/Moses/mosesdecoder/bin/moses train/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ --decoder-flags="-threads 6" &> mert.out &
[1] 12508
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ top

top - 12:29:17 up 2:09, 2 users, load average: 3.58, 1.41, 0.62
Tasks: 153 total, 1 running, 152 sleeping, 0 stopped, 0 zombie
Cpu(s): 95.7%us, 4.3%sy, 0.0%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 6786756k total, 4620088k used, 2166668k free, 215040k buffers
Swap: 7336956k total, 0k used, 7336956k free, 2335844k cached

PID USER      PR  NI    VIRT    RES    SHR S %CPU %MEM     TIME+ COMMAND
12512 nlp2ct    20   0 1102m  800m   31m S 85.8 12.1    0:48.32 moses
23689 nlp2ct    20   0 1252m  403m   53m S  5.0  6.1    2:08.16 firefox
```

Type “ctrl+c”, “ps -aux” to show detailed running,

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ps -aux
Warning: bad ps syntax, perhaps a bogus '-'? See http://procps.sf.net/faq.html
USER      PID %CPU %MEM      VSZ      RSS TTY      STAT START   TIME COMMAND
root        1  0.0  0.0  24456  2364 ?          Ss   10:19   0:00 /sbin/init
root        2  0.0  0.0      0      0 ?          S   10:19   0:00 [kthreadd]
root        3  0.0  0.0      0      0 ?          S   10:19   0:01 [ksoftirqd/0]
root        5  0.0  0.0      0      0 ?          S   10:19   0:00 [kworker/u:0]
root        6  0.0  0.0      0      0 ?          S   10:19   0:00 [migration/0]
root        7  0.0  0.0      0      0 ?          S   10:19   0:00 [watchdog/0]
```

Type “ps –aux|grep moses” to show the running that only related to moses:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ps -aux|grep moses
Warning: bad ps syntax, perhaps a bogus '-'? See http://procps.sf.net/faq.html
nlp2ct  2182  0.4  1.3 587756 93516 ?          Sl   10:34   0:33 gedit /home/nlp2ct/Aaron/working/train/model/moses.ini
nlp2ct  12508  0.0  0.0  29720  5532 pts/3      S   12:28   0:00 /usr/bin/perl -w /home/nlp2ct/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl /home/nlp2ct/Aaron/corpus/tune/news-test2008.true.fr /home/nlp2ct/Aaron/corpus/tune/news-test2008.true.en /home/nlp2ct/Aaron/Moses/mosesdecoder/bin/moses train/model/moses.ini --mertdir /home/nlp2ct/Aaron/Moses/mosesdecoder/bin/ --decoder-flags=-threads 6
nlp2ct  12511  0.0  0.0  4404   608 pts/3      S   12:28   0:00 sh -c /home/nlp
```

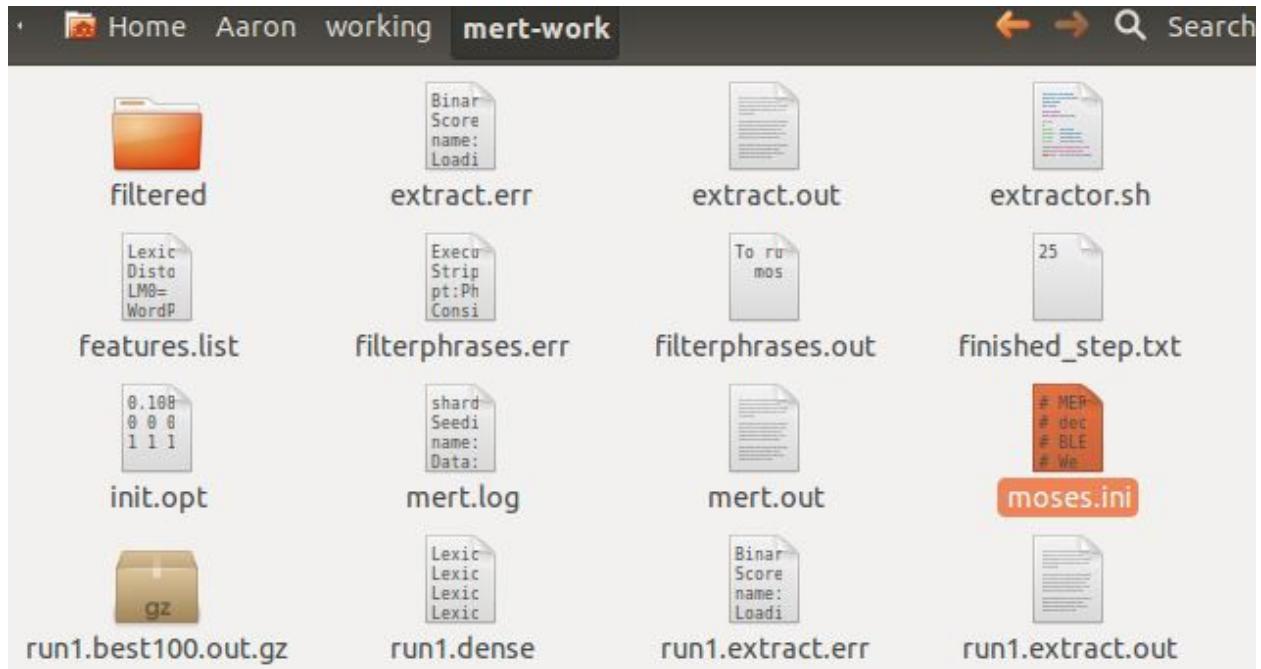
A long time later, type “top” showing there will be no running of “moses”, which means the tuning is finished:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ top

top - 10:06:37 up 23:47, 2 users, load average: 3.15, 1.45, 1.44
Tasks: 152 total, 2 running, 150 sleeping, 0 stopped, 0 zombie
top - 10:53:09 up 1 day, 33 min, 2 users, load average: 0.09, 0.10, 0.42
Tasks: 152 total, 3 running, 149 sleeping, 0 stopped, 0 zombie
Cpu(s): 1.0%us, 0.7%sy, 0.0%ni, 98.3%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 6786756k total, 6664780k used, 121976k free, 6220k buffers
Swap: 7336956k total, 2599052k used, 4737904k free, 77656k cached

PID USER PR NI VIRT RES SHR S %CPU %MEM TIME+ COMMAND
12260 nlp2ct 20 0 396m 11m 2628 R 1.3 0.2 24:41.10 plugin-containe
1142 root 20 0 369m 171m 9420 S 0.3 2.6 6:24.51 Xorg
16606 nlp2ct 20 0 17340 800 436 R 0.3 0.0 0:07.77 top
```

There will be one document “mert.out” generated under working file, at the same time, a “mert-work” file containing **the tuning results “moses.ini”** is generated under working file as below:



Begin the tuning at 2013-10-29-17:00, finish the tuning and generate the “moses.ini” file at 2013-10-30-09:44, it takes **around 16 hours** (using 7 threads). The “moses.ini” file contains the tuned parameters and other information as following:

```

# MERT optimized configuration
# decoder /home/nlp2ct/Aaron/Moses/mosesdecoder/bin/moses
# BLEU 0.186923 on dev /home/nlp2ct/Aaron/corpus/tune/news-test2008.true.fr
# We were before running iteration 25
# finished Wed Oct 30 09:44:37 CST 2013
## MOSES CONFIG FILE ##
#####
##### input factors #####
#[input-factors]
0

##### mapping steps #####
#[mapping]
0 T 0

#[distortion-limit]
6

##### feature functions #####
#[feature]
UnknownWordPenalty
WordPenalty
PhrasePenalty
PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/nlp2ct/Aaron/working/train/model/phrase-table.gz input-factor=0 output-factor=0
LexicalReordering name=LexicalReordering0 num-features=6 type=wbe-msd-bidirectional-fe-allff input-factor=0 output-factor=0 path=/home/nlp2ct/Aaron/working/train/model/reordering-table.wbe-msd-bidirectional-fe.gz
Distortion
KENLM lazyken=0 name=LM0 factor=0 path=/home/nlp2ct/Aaron/lm/news-commentary-v8.fr-en.blm.en order=3

##### dense weights for feature functions #####
#[threads]
7
#[weight]

LexicalReordering0= 0.108647 0.0221523 0.0679283 0.139558 0.0377834 0.0857802
Distortion0= 0.0713619
LM0= 0.0763578
WordPenalty0= -0.0298535
PhrasePenalty0= 0.124991
TranslationModel0= 0.0275796 0.0760543 0.0925488 0.0394041

```

In the file “working/mert.out”, it shows:

```
(25) BEST at 25: 0.108647 0.0221523 0.0679283 0.139558 0.0377834 0.0857802 0.0713
Executing: \cp -f mert.log run25.mert.log
featlist: LexicalReordering0=0.108647
featlist: LexicalReordering0=0.0221523
featlist: LexicalReordering0=0.0679283
featlist: LexicalReordering0=0.139558
featlist: LexicalReordering0=0.0377834
featlist: LexicalReordering0=0.0857802
featlist: Distortion0=0.0713619
featlist: LM0=0.0763578
featlist: WordPenalty0=-0.0298535
featlist: PhrasePenalty0=0.124991
featlist: TranslationModel0=0.0275796
featlist: TranslationModel0=0.0760543
featlist: TranslationModel0=0.0925488
featlist: TranslationModel0=0.0394041
Parsing --decoder-flags: |-threads 7|
Saving new config to: ./moses.ini
Saved: ./moses.ini
Training finished at Wed Oct 30 09:44:37 CST 2013
```

```
=====
```

Testing the translation model:

This stage is to test the translation quality of the built translation model. The translation quality is measured by the automatic evaluation metric score, such as the metric BLEU(2002), METEOR(2005), and LEOPR(2012), etc.

Type the command:

```
cd ~/Aaron
~/Aaron/Moses/mosesdecoder/bin/moses -f ~/Aaron/working/mert-work/moses.ini
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron$ ~/Aaron/Moses/mosesdecoder/bin/moses -f ~/Aaron/working/mert-work/moses.ini
Defined parameters (per moses.ini or switch):
    config: /home/nlp2ct/Aaron/working/mert-work/moses.ini
    distortion-limit: 6
    feature: UnknownWordPenalty WordPenalty PhrasePenalty PhraseDictionaryMe
```

It will take some minutes to show the finish the initializing LexicalReordering and reading phrase-table as following:

```
Initializing LexicalReordering..
line=Distortion
FeatureFunction: Distortion0 start: 13 end: 13
line=KENLM lazyken=0 name=LMO factor=0 path=/home/nlp2ct/Aaron/lm/news-commentar
y-v8.fr-en.blm.en order=3
FeatureFunction: LMO start: 14 end: 14
Loading table into memory...done.
Start loading text SCFG phrase table. Moses format : [128.000] seconds
Reading /home/nlp2ct/Aaron/working/train/model/phrase-table.gz
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80
---85---90---95--100
*****
IO from STDOUT/STDIN
Created input-output object : [996.000] seconds
```

Type a French sentence “c'est une petite maison.(//this is a small house.)” and “enter”, it will translate it as below, which is not a fully translated sentence:

```
Created input-output object : [996.000] seconds
c'est une petite maison.
Translating line 0  in thread id 140224420796160
Translating: c'est une petite maison.
Line 0: Collecting options took 1.000 seconds
Line 0: Search took 4.000 seconds
c'est a small maison.
BEST TRANSLATION: c'est|UNK|UNK|UNK a small maison.|UNK|UNK|UNK [1111]  [total=-202.004] core=(-200.000,-4.000,3.000,-1.392,-3.215,-0.705,-1.600,-0.403,0.000,0.000,-0.181,0.000,0.000,0.000,-26.435)
Line 0: Translation took 5.000 seconds total
```

Type another French sentence “vous êtes beau (// you are handsome)”, it will translate it into “you are beautiful” as below:

```
vous êtes beau
Translating line 1  in thread id 140224404010752
Translating: vous êtes beau
Line 1: Collecting options took 1.000 seconds
Line 1: Search took 0.000 seconds
you are beautiful
BEST TRANSLATION: you are beautiful [111]  [total=-2.952] core=(0.000,-3.000,2.00,-3.737,-7.607,-2.750,-4.150,-0.538,0.000,0.000,-0.280,0.000,0.000,0.000,-27.434)
Line 1: Translation took 1.000 seconds total
```

Type “ctrl+c” to exit the job.

To make the translation faster, we should binary the lexicalized reordering models and the phrase-table.

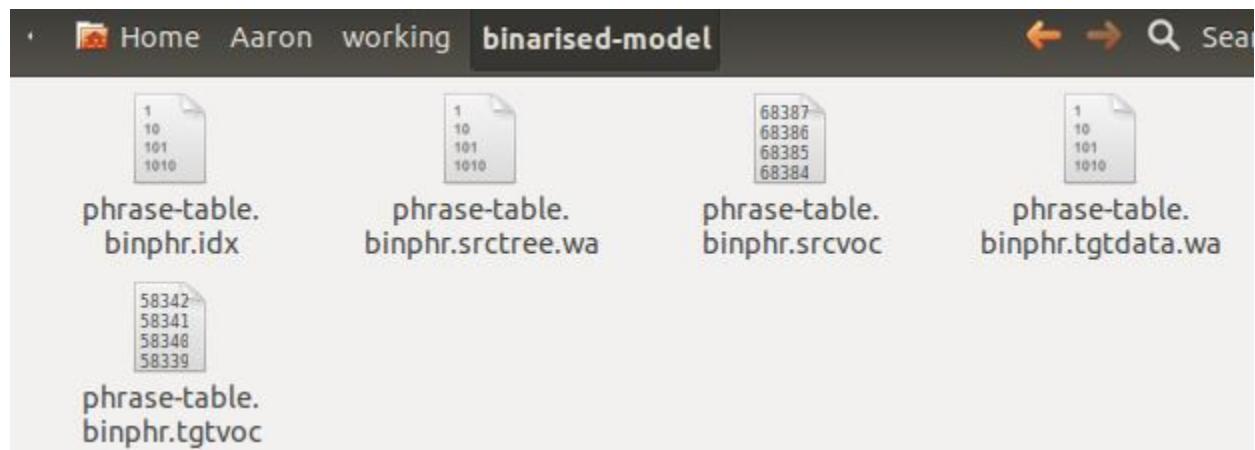
```
mkdir ~/Aaron/working/binarised-model
```

Type the following command to binarise the phrase-table:

```
~/Aaron/Moses/mosesdecoder/bin/processPhraseTable -ttable 0 0 train/model/phrase-table.gz  
-nscores 5 -out binarised-model/phrase-table
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/processPhraseTable -ttable 0 0 train/model/phrase-table.gz -nscores 5 -out binarised-model/phrase-table  
processing ptree for train/model/phrase-table.gz  
..... [phrase:500000]  
..... [phrase:1000000]  
..... [phrase:1500000]  
..... [phrase:2000000]  
..... [phrase:2500000]  
..... [phrase:3000000]  
..... [phrase:3500000]  
..... [phrase:4000000]  
..... [phrase:4500000]  
..... [phrase:5000000]  
..... [phrase:5500000]  
..... [phrase:6000000]  
..... [phrase:6500000]  
..... [phrase:7000000]  
..... [phrase:7500000]  
...distinct source phrases: 7537247 distinct first words of source phrases: 6838  
8 number of phrase pairs (line count): 11424565  
Count of lines with missing alignments: 0/11424565  
WARNING: there are src voc entries with no phrase translation: count 9684  
There exists phrase translations for 58704 entries  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

It will generate the following files:



To binary the lexical reordering-table, type the following command:

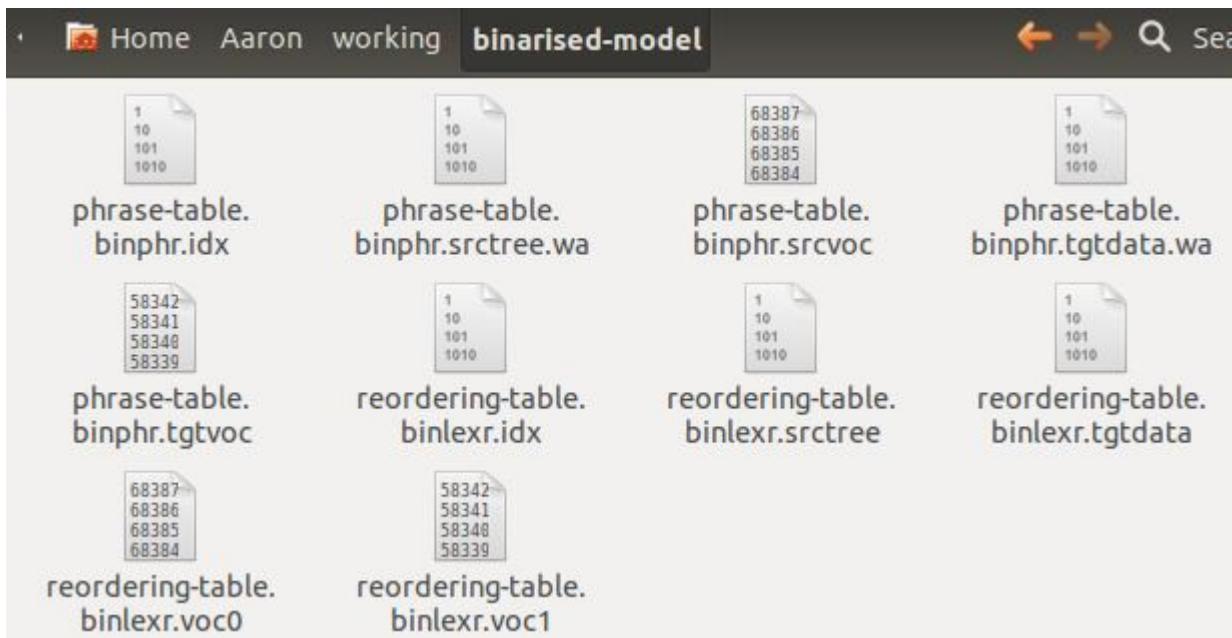
```
~/Aaron/Moses/mosesdecoder/bin/processLexicalTable -in train/model/reordering-table.wbe-msd-bidirectional-fe.gz -out binarised-model/reordering-table
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/processLexicalTable -in train/model/reordering-table.wbe-msd-bidirectional-fe.gz -out binarised-model/reordering-table
processLexicalTable v0.1 by Konrad Rawlik
processing train/model/reordering-table.wbe-msd-bidirectional-fe.gz to binarised-model/reordering-table.*
```

.....

```
..... nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

It will generate the files:



Copy the ~/working/mert-work/moses.ini document into the ~/binarised-model. Then change the moses.ini content as below to point to the binarised files:

```

# feature functions
[feature]
UnknownWordPenalty
WordPenalty
PhrasePenalty
PhraseDictionaryBinary name=TranslationModel0 table-limit=20 num-features=4 path=/home/nlp2ct/Aaron/
working/binarised-model/phrase-table input-factor=0 output-factor=0
LexicalReordering name=LexicalReordering0 num-features=6 type=wbe-msd-bidirectional-fe-allff input-
factor=0 output-factor=0 path=/home/nlp2ct/Aaron/working/binarised-model/reordering-table
Distortion
KENLM lazyken=0 name=LM0 factor=0 path=/home/nlp2ct/Aaron/lm/news-commentary-v8.fr-en.blm.en order=3

```

Using the binarised tables, the loading will be very fast using the following command:
`/Aaron/Moses/mosesdecoder/bin/moses -f ~/Aaron/working/binarised-model/moses.ini

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses -f ~/Aaron/working/binarised-model/moses.ini
```

If you type the above two french sentences again it will generate the same translations as before.

Type the commands to prepare the testing data as below:

```

mkdir ~/Aaron/corpus/test
cd ~/corpus/test
~/Aaron/Moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en <~/Aaron/corpus/tune/dev/newstest2011.en > newstest2011.tok.en

~/Aaron/Moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l fr <~/Aaron/corpus/tune/dev/newstest2011.fr > newstest2011.tok.fr

~/Aaron/Moses/mosesdecoder/scripts/recaser/truecase.perl --model ~/Aaron/corpus/truecase-
model.en < newstest2011.tok.en > newstest2011.true.en

~/Aaron/Moses/mosesdecoder/scripts/recaser/truecase.perl --model ~/Aaron/corpus/truecase-
model.fr < newstest2011.tok.fr > newstest2011.true.fr

```

To make the translation faster, we can filter the trained translation model to retain the entries that are only needed to translate the offered test corpus:

```

VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/filter-model-
given-input.pl filtered-newstest2011 mert-work/moses.ini
~/Aaron/corpus/test/newstest2011.true.fr -Binarizer
~/Aaron/Moses/mosesdecoder/bin/processPhraseTable

```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/filter-model-given-input.pl filtered-newstest2011 mert-work/moses.ini ~/Aaron/corpus/test/newstest2011.true.fr -Binarizer ~/Aaron/Moses/mosesdecoder/bin/processPhraseTable
```

The filtering and binary stage finished as below:

```
filtering /home/nlp2ct/Aaron/working/train/model/reordering-table.wbe-msd-bidirectional-fe.gz -> /home/nlp2ct/Aaron/working/filtered-newstest2011/reordering-table.wbe-msd-bidirectional-fe...
1054277 of 11424565 phrases pairs used (9.23%) - note: max length 10
binarizing.../home/nlp2ct/Aaron/Moses/mosesdecoder/bin/processLexicalTable -in /home/nlp2ct/Aaron/working/filtered-newstest2011/reordering-table.wbe-msd-bidirectional-fe.gz -out /home/nlp2ct/Aaron/working/filtered-newstest2011/reordering-table.wbe-msd-bidirectional-fe
processLexicalTable v0.1 by Konrad Rawlik
processing /home/nlp2ct/Aaron/working/filtered-newstest2011/reordering-table.wbe-msd-bidirectional-fe.gz to /home/nlp2ct/Aaron/working/filtered-newstest2011/reordering-table.wbe-msd-bidirectional-fe.*
```

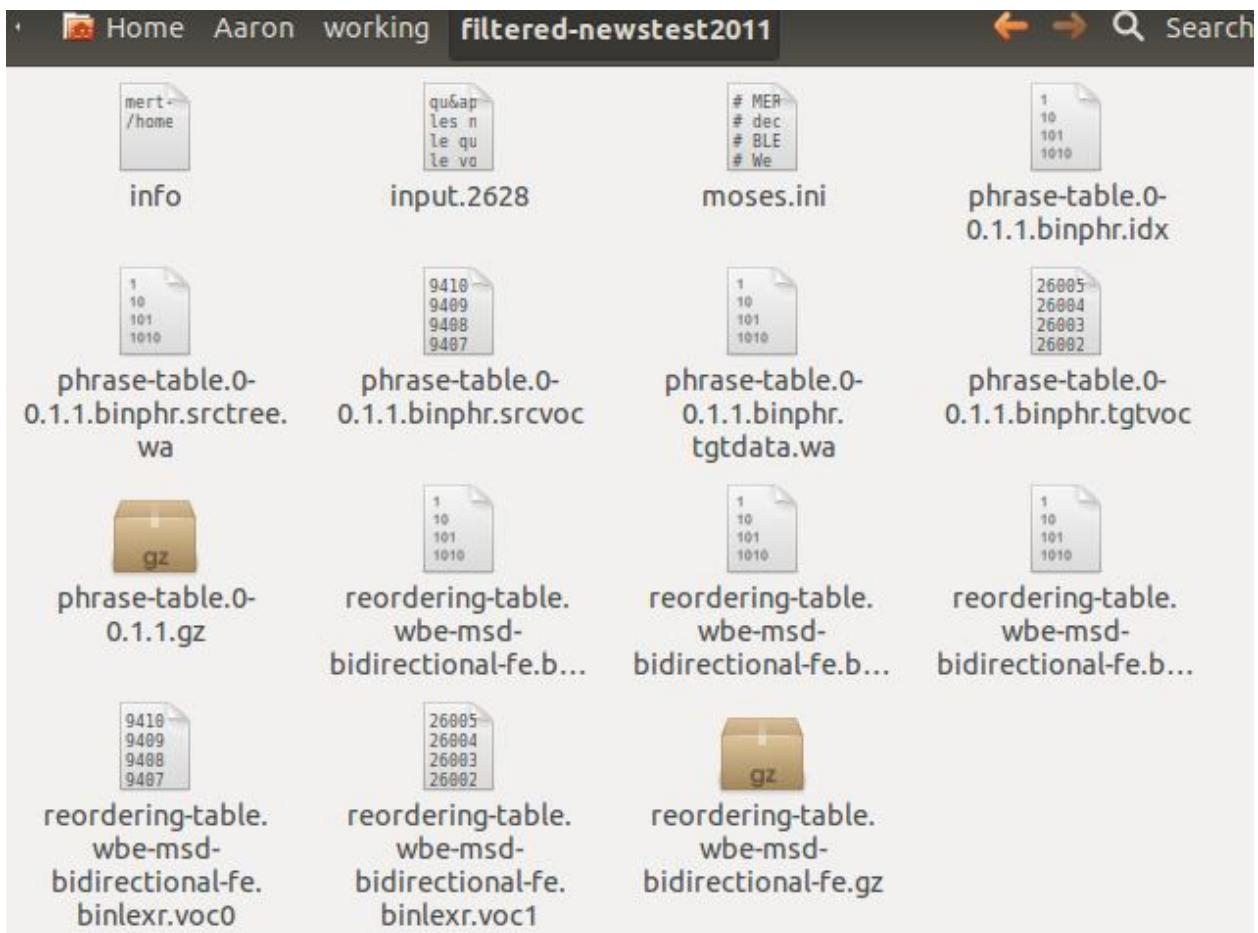
.....

.....To run the decoder, please call:

```
moses -f /home/nlp2ct/Aaron/working/filtered-newstest2011/moses.ini -i /home/nlp2ct/Aaron/working/filtered-newstest2011/input.2628
```

nlp2ct@nlp2ct-VirtualBox:~/Aaron/working\$

Above command will generate a “~/Aaron/working/filtered-newstest2011” file that contains the following documents:



Then we translate the testing corpus and score the translation quality using BLEU metric:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses -f ~/Aaron/working/filtered-newstest2011/moses.ini < ~/Aaron/corpus/test/newstest2011.true.fr > ~/Aaron/working/newstest2011.translated.en 2> ~/Aaron/working/newstest2011.out
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses -f ~/Aaron/working/filtered-newstest2011/moses.ini < ~/Aaron/corpus/test/newstest2011.true.fr > ~/Aaron/working/newstest2011.translated.en 2> ~/Aaron/working/newstest2011.out
```

It takes about 25 minutes to finish the translation. There will be two documents generated “~/working/newstest2011.out & ~/working/newstest2011.translated.en”. The document newstest2011.translated.en contains the translated output sentences; the document newstest2011.out contains the detailed translation procedure and its volume is larger:

```
newstest2011.translated.en newstest2011.out
that is what will happen ? CSSD is neither Voldemort or pâtisserie in Prague .
the new CSSD elected to the mayor of the capital will apparently overcome a real linguistic barrier to
understand the level of city council with their old and new colleagues of the ODS .
the newspaper Actualne.cz has " tested " the new members of the city council belonging to
CSSD , whether they know the argot who was born over the years dernieres mayoral , when it were the
current coalition partners that ruled in Prague .
the vocabulary codé that the best policy of Prague put underway over the previous era of the mayor
Pavel Bem described a few people , situations , and causes of the mainland .
it is with surprise that he has been discovered that the new elected do not understand très well the
terms known .
at least the they say .
" that is Voldemort . "
```

```
newstest2011.translated.en newstest2011.out
defined parameters (per moses.ini or switch):
    config: /home/nlp2ct/Aaron/working/filtered-newstest2011/moses.ini
    distortion-limit: 6
    feature: UnknownWordPenalty WordPenalty PhrasePenalty PhraseDictionaryBinary
name=TranslationModel0 table-limit=20 num-features=4 path=/home/nlp2ct/Aaron/working/filtered-
newstest2011/phrase-table.0-0.1.1 input-factor=0 output-factor=0 LexicalReordering
name=LexicalReordering0 num-features=6 type=wbe-msd-bidirectional-fe-allff input-factor=0 output-
factor=0 path=/home/nlp2ct/Aaron/working/filtered-newstest2011/reordering-table.wbe-msd-bidirectional-
fe Distortion KENLM lazyken=0 name=LM0 factor=0 path=/home/nlp2ct/Aaron/lm/news-commentary-v8.fr-
en.blm.en order=3
    input-factors: 0
    mapping: 0 T 0
    threads: 7
    weight: LexicalReordering0= 0.108647 0.0221523 0.0679283 0.139558 0.0377834 0.0857802
Distortion0= 0.0713619 LM0= 0.0763578 WordPenalty0= -0.0298535 PhrasePenalty0= 0.124991
TranslationModel0= 0.0275796 0.0760543 0.0925488 0.0394041
/home/nlp2ct/Aaron/Moses/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryBinary name=TranslationModel0 table-limit=20 num-features=4 path=/home/nlp2ct/
```

Type the following command to test the BLEU score of the automatic translation, as compared with the reference translation:

```
VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/generic/multi-bleu.perl -lc
~/Aaron/corpus/test/newstest2011.true.en < ~/Aaron/working/newstest2011.translated.en
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/gen-
eric/multi-bleu.perl -lc ~/Aaron/corpus/test/newstest2011.true.en < ~/Aaron/work-
ing/newstest2011.translated.en
BLEU = 23.41, 60.0/29.8/16.8/10.0 (BP=1.000, ratio=1.017, hyp_len=76012, ref_len=
74753)
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

It shows that the BLEU score is 23.41.

```
=====
```

Trial of Tree to tree translation using Moses:

Install the chart decoder:

We already install the chart decoder as the executable file:

~/Aaron/Moses/mosesdecoder/bin/moses_chart

First use the following command to try the string to tree model:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/sample-models$ ~/  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/sample-models$ ~/Aaron/Moses/mosesdecoder  
/bin/moses_chart -f string-to-tree/moses.ini < string-to-tree/in >out.stt
```

The running will call the language model: ~/lm/europarl.srilm.gz

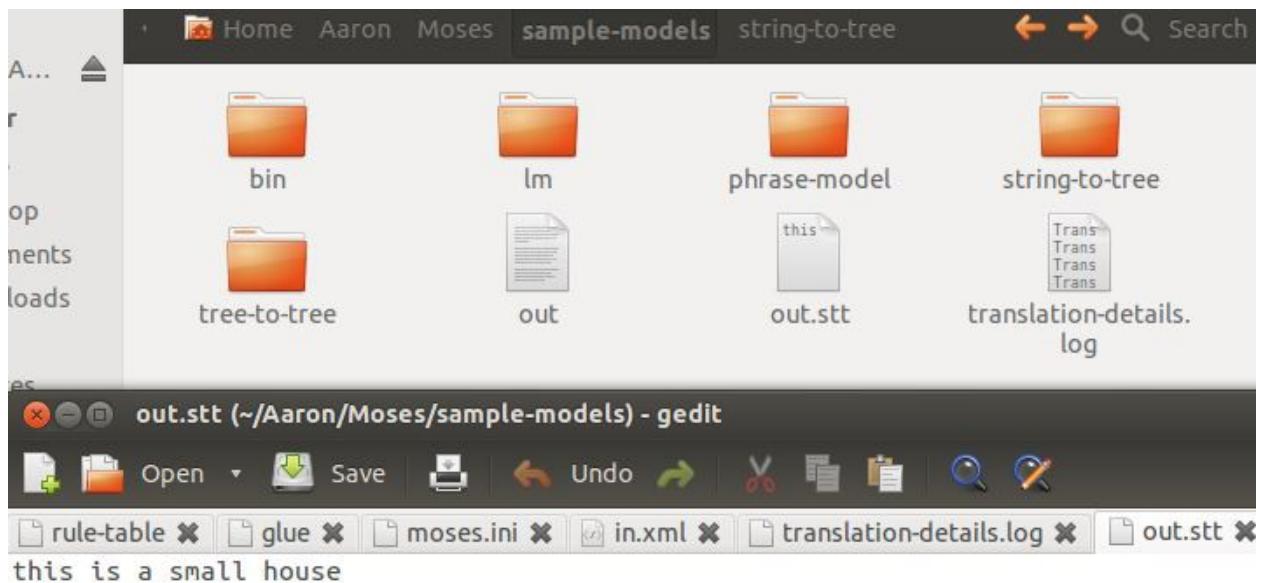
```
Defined parameters (per moses.ini or switch):
    config: string-to-tree/moses.ini
    cube-pruning-pop-limit: 1000
    feature: KENLM name=LM factor=0 order=3 num-features=1 path=lm/europarl.
srilm.gz WordPenalty UnknownWordPenalty PhraseDictionaryMemory input-factor=0 ou
tput-factor=0 path=string-to-tree/rule-table num-features=1 table-limit=20
    input-factors: 0
    inputtype: 3
    mapping: 0 T 0
    max-chart-span: 20 1000
    non-terminals: X S
    search-algorithm: 3
    translation-details: translation-details.log
    weight: WordPenalty0= 0 LM= 0.5 PhraseDictionaryMemory0= 0.5
/home/nlp2ct/Aaron/Moses/mosesdecoder/bin
line=KENLM name=LM factor=0 order=3 num-features=1 path=lm/europarl.srilm.gz
FeatureFunction: LM start: 0 end: 0
Loading the LM will be faster if you build a binary file.
Reading lm/europarl.srilm.gz
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80
---85---90---95--100
**The ARPA file is missing <unk>. Substituting log10 probability -100.000.
*****
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 2 end: 2
line=PhraseDictionaryMemory input-factor=0 output-factor=0 path=string-to-tree/r
ule-table num-features=1 table-limit=20
FeatureFunction: PhraseDictionaryMemory0 start: 3 end: 3
Start loading text SCFG phrase table. Moses format : [1.000] seconds
Reading string-to-tree/rule-table
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80
---85---90---95--100
*****
```

```
*****
max-chart-span: 20
IO from STDOUT/STDIN
Created input-output object : [1.000] seconds
Translating: <s> das ist ein kleines haus </s> ||| [0,0]=X (1) [0,1]=X (1) [0,2]
=X (1) [0,3]=X (1) [0,4]=X (1) [0,5]=X (1) [0,6]=X (1) [1,1]=X (1) [1,2]=X (1)
[1,3]=X (1) [1,4]=X (1) [1,5]=X (1) [1,6]=X (1) [2,2]=X (1) [2,3]=X (1) [2,4]=X
(1) [2,5]=X (1) [2,6]=X (1) [3,3]=X (1) [3,4]=X (1) [3,5]=X (1) [3,6]=X (1) [4,4]
=X (1) [4,5]=X (1) [4,6]=X (1) [5,5]=X (1) [5,6]=X (1) [6,6]=X (1)

    0   1   2   3   4   5   6
    0   3   2   2   2   1   0
    0   0   0   0   0   0   0
    0   0   0   3   0
    0   0   4   0
    0   4   0
    0   0
    1

BEST TRANSLATION: 41 TOP -> <s> S </s> :1-1 : c=-3.206 core=(-6.413,-2.000,0.00
0,0.000) [0..6] 20 [total=-15.501] core=(-27.091,-7.000,0.000,-3.912)
Translation took 0.000 seconds
End. : [1.000] seconds
Name:moses_chart          VmPeak:190440 kB          VmRSS:31860 kB  RSSMax:32428 kB
ser:0.656      sys:0.092      CPU:0.748      real:0.840
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/sample-models$
```

It will generate the document “out.stt” and “translation-details.log” under the directory:
 ~/Aaron/Moses/sample-models



```

translation-details.log (~/Aaron/Moses/sample-models) - gedit
File Open Save Undo Redo Cut Copy Paste Find Replace
rule-table glue moses.ini in.xml translation-details.log out.stt
Trans Opt 0 [0..6]: [6..6]=</s> [1..5]=S [0..0]=<s> : TOP ->TOP -> <s> S </s> :1-1 : c=-3.20649
core=(-6.41298,-2,0,0) -15.5014core=(-27.0908,-7,0,-3.91202)
Trans Opt 0 [1..5]: [2..5]=VP [1..1]=NP : S ->S -> NP VP :0-0 1-1 : c=0 core=(0,-0,0,0)
-14.9803core=(-26.0485,-5,0,-3.91202)
Trans Opt 0 [1..1]: [1..1]=das : NP ->NP -> this :: c=-4.00102 core=(0,-1,0,-2.30259) -4.00102core=
(-5.69946,-1,0,-2.30259)
Trans Opt 0 [2..5]: [3..5]=NP [2..2]=V : VP ->VP -> V NP :0-0 1-1 : c=0 core=(0,-0,0,0)
-12.0877core=(-22.566,-4,0,-1.60944)
Trans Opt 0 [2..2]: [2..2]=ist : V ->V -> is :: c=-2.46111 core=(0,-1,0,0) -2.46111core=
(-4.92223,-1,0,0)
Trans Opt 0 [3..5]: [5..5]=NN [4..4]=ADJ [3..3]=DT : NP ->NP -> DT ADJ NN :0-0 1-1 2-2 : c=0 core=
(0,-0,0,0) -10.9498core=(-20.2901,-3,0,-1.60944)
Trans Opt 0 [3..3]: [3..3]=ein : DT ->DT -> a :: c=-2.75755 core=(0,-1,0,0) -2.75755core=
(-5.5151,-1,0,0)
Trans Opt 0 [4..4]: [4..4]=kleines : ADJ ->ADJ -> small :: c=-4.86058 core=(0,-1,0,-1.60944)
-4.86058core=(-8.11173,-1,0,-1.60944)
Trans Opt 0 [5..5]: [5..5]=haus : NN ->NN -> house :: c=-4.63303 core=(0,-1,0,0) -4.63303core=
(-9.26607,-1,0,0)

```

Secondly use the following command to try the tree-to-tree model:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/Moses/sample-models$ ~/Aaron/Moses/mosesdecoder
/bin/moses_chart -f tree-to-tree/moses.ini < tree-to-tree/in.xml > out.ttt
```

The in.xml file contains the sentence:

```

<tree label="TOP"><tree label="OS">overhead</tree> <tree label="NS">oxygen</tree> <tree label="NS">masks</tree> in the <tree
label="NS">cabin <tree label="NS">section</tree></tree> <tree label="OS">had dropped into place .</tree></tree>

```

The running calls the rule table: ~/tree-to-tree/rule-table

```
rule-table ✘ glue ✘ moses.ini ✘ In.xml ✘ translation-details.log ✘ out.ttt ✘
overhead [OS] ||| 头顶上的 [OT] ||| 0.419355 0.380952 1 0.8 2.718 ||| ||| 31 13
masks [NS] ||| 面罩 [NT] ||| 1 0.416667 1 0.263158 2.718 ||| ||| 1 1
oxygen [NS] ||| 氧气 [NT] ||| 1 0.416667 1 0.263158 2.718 ||| ||| 1 1
[NS][NT] in the [NS][NT] [NS] ||| [NS][NT] 的 [NS][NT] [NT] ||| 0.227273 0.0248669 1 0.0484183 2.718 ||| 3-0 0-2 ||| 4.03333 0.916666
cabin section [NS] ||| 船舱 区 [NT] ||| 1 0.416667 1 0.263158 2.718 ||| ||| 1 1
section [NS] ||| 区 [NT] ||| 1 0.416667 1 0.263158 2.718 ||| ||| 1 1
had dropped into place . [OS] ||| 已滑落 。 [OT] ||| 0.00694444 0.00265754 1 0.170034 2.718 ||| ||| 1 1
[NS][NT] [NS][NT] [NS] ||| [NS][NT] [NS][NT] [NT] ||| 1 1 1 1 1 ||| 0-0 1-1 ||| 1 1
[NS][NT] [NS][NT] in the [NS][NT] [NS] ||| [NS][NT] [NS][NT] [NS][NT] [NT] ||| 1 1 1 1 1 ||| 0-0 1-1 4-2 ||| 1 1
```

```

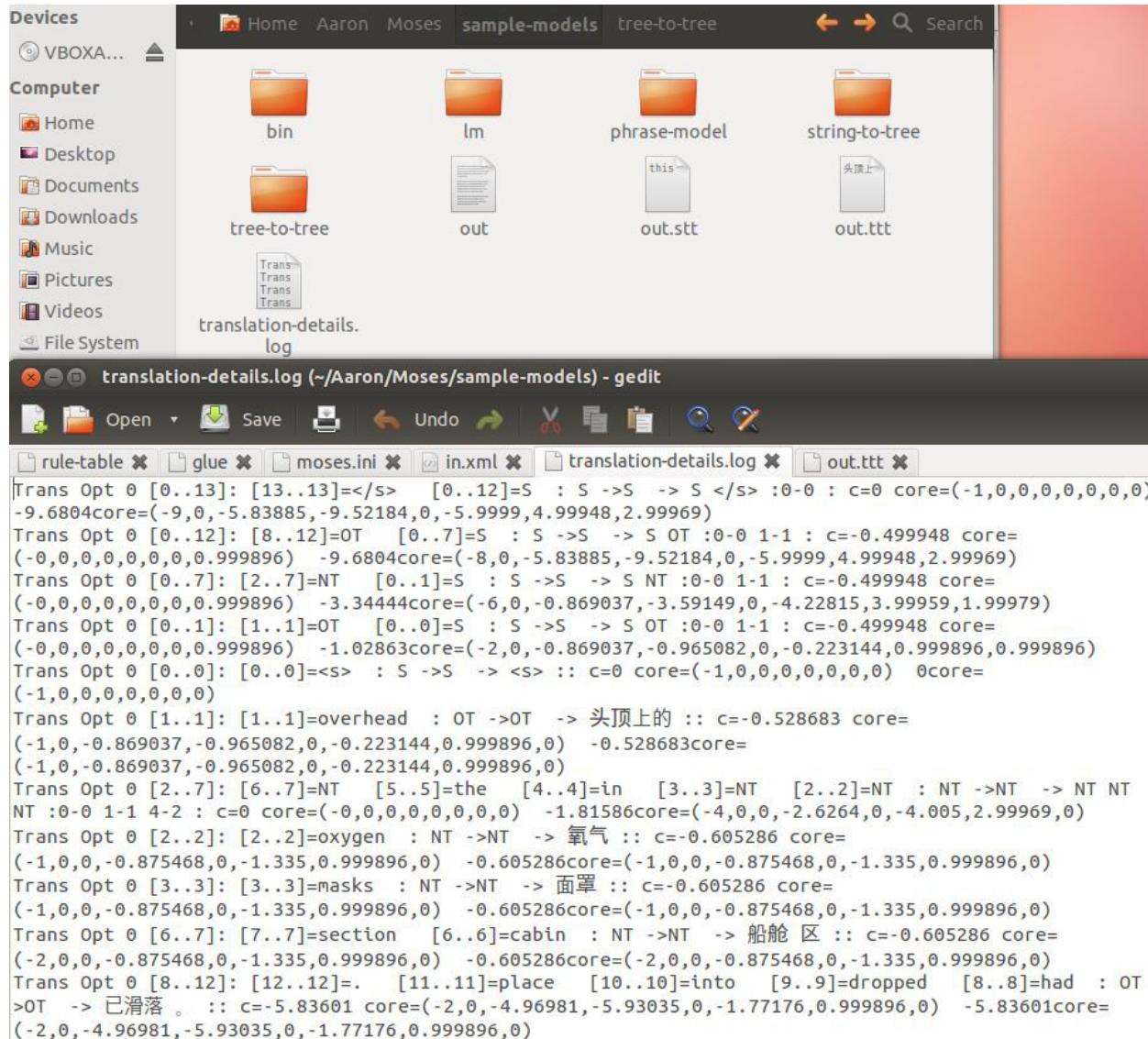
Defined parameters (per moses.ini or switch):
    config: tree-to-tree/moses.ini
    cube-pruning-pop-limit: 1000
    feature: WordPenalty UnknownWordPenalty PhraseDictionaryMemory input-factor=0 output-factor=0 path=tree-to-tree/rule-table num-features=5 table-limit=20
    PhraseDictionaryMemory input-factor=0 output-factor=0 path=tree-to-tree/glue num-features=1 table-limit=20
    input-factors: 0
    inputtype: 3
    mapping: 0 T 0 1 T 1
    max-chart-span: 20 1000
    non-terminals: X S
    search-algorithm: 3
    translation-details: translation-details.log
    weight: WordPenalty0= 0 PhraseDictionaryMemory0= 0.5 0.5 0.5 0.5 0.5 PhraseDictionaryMemory1= -0.5
/home/nlp2ct/Aaron/Moses/mosesdecoder/bin
line=WordPenalty
FeatureFunction: WordPenalty0 start: 0 end: 0
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 1 end: 1
line=PhraseDictionaryMemory input-factor=0 output-factor=0 path=tree-to-tree/rule-table num-features=5 table-limit=20
FeatureFunction: PhraseDictionaryMemory0 start: 2 end: 6
line=PhraseDictionaryMemory input-factor=0 output-factor=0 path=tree-to-tree/glue num-features=1 table-limit=20
FeatureFunction: PhraseDictionaryMemory1 start: 7 end: 7
Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading tree-to-tree/rule-table
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80
---85---90---95---100
*****
***** Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading tree-to-tree/glue
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80
---85---90---95---100
*****

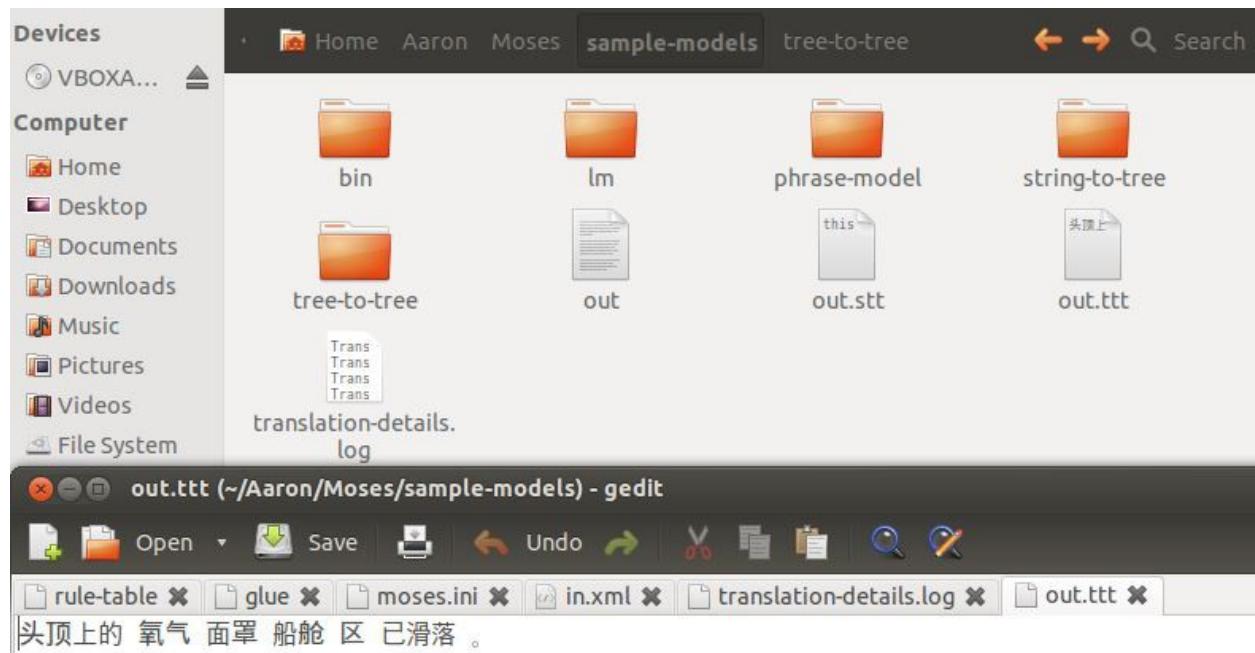
```

```

*****
max-chart-span: 20
max-chart-span: 1000
IO from STDOUT/STDIN
Created input-output object : [0.000] seconds
Translating: <s> overhead oxygen masks in the cabin section had dropped into place . </s> ||| [0,0]=X (1) [0,1]=X (1) [0,2]=X (1) [0,3]=X (1) [0,4]=X (1) [0,5]=X (1) [0,6]=X (1) [0,7]=X (1) [0,8]=X (1) [0,9]=X (1) [0,10]=X (1) [0,11]=X (1) [0,12]=X (1) [0,13]=X (1) [1,1]=X (1) [1,1]=OS (1) [1,2]=X (1) [1,3]=X (1) [1,4]=X (1) [1,5]=X (1) [1,6]=X (1) [1,7]=X (1) [1,8]=X (1) [1,9]=X (1) [1,10]=X (1) [1,11]=X (1) [1,12]=TOP (1) [1,13]=X (1) [2,2]=X (1) [2,2]=NS (1) [2,3]=X (1) [2,4]=X (1) [2,5]=X (1) [2,6]=X (1) [2,7]=X (1) [2,8]=X (1) [2,9]=X (1) [2,10]=X (1) [2,11]=X (1) [2,12]=X (1) [2,13]=X (1) [3,3]=X (1) [3,3]=NS (1) [3,4]=X (1) [3,5]=X (1) [3,6]=X (1) [3,7]=X (1) [3,8]=X (1) [3,9]=X (1) [3,10]=X (1) [3,11]=X (1) [3,12]=X (1) [3,13]=X (1) [4,4]=X (1) [4,5]=X (1) [4,6]=X (1) [4,7]=X (1) [4,8]=X (1) [4,9]=X (1) [4,10]=X (1) [4,11]=X (1) [4,12]=X (1) [4,13]=X (1) [5,5]=X (1) [5,6]=X (1) [5,7]=X (1) [5,8]=X (1) [5,9]=X (1) [5,10]=X (1)
```

When finishing it generates document “out.ttt” and update the document “translation-details.log”





Build ZH->EN Tree to tree translation using Moses:

Prepare corpus:

The corpus “Oxford-dictionary” is extracted from Oxford Chinese-English dictionary (7th version), and segmented using NLPIR2013 segmentor (Zhang Huangping, CAS, accuracy 0.97).

Using the Oxford-dictionary.zh 12000 tokenized simplified Chinese sentences, the GrammarTrained on CTB-7, BerkeleyParser-1.7.jar, the parsing information: 10 minutes.

Using the TrainedGramEng-WSJ, Berkeley-Parser-1.7.jar, Oxford-dictionary.zh 12000 tokenized English sentences, the parsing: 11 minutes.

Firstly, we select the first 1200 sentences bilingual EN-ZH corpora for a trial. The parsed corpora as below:

	Oxford-dic1200.en.parsed	Oxford-dic1200.en.parsed	Oxford-dic1200.zh.parsed
1	((S (NP (NP (DT The) (JJ first) (NN letter)) (PP (IN in) (NP (DT the) (NNP English) (NNP Alphabet)))) (VP (VBZ is) (NP (NNP A))) (. . .)))	是	是
2	((S (NP (PRP He)) (VP (VBD earned) (NP (CD four) (NP (NNP A) (POS 's)) (PP (IN in) (NP (DT the) (JJ final) (NNS examinations)))) (. . .)))	是	是
3	((S (NP (PRP He)) (VP (VBZ has) (VP (VBN got) (NP (NP (DT a) (NN job)) (PP (IN in) (NP (NNP Los) (NNP Angeles)))))) (. . .)))	是	是
4	((S (S (NP (PRP I)) (VP (VBP have) (NP (DT a) (JJ bad) (JJ cold)))) (CC and) (S (NP (PRP I)) (VP (VBP am) (ADJP (JJ full) (PP (IN of) (NP (NN	是	是
5	((S (S (NP (PRP He)) (VP (VBD was) (ADVP (RB out) (PP (IN in) (NP (DT the) (NN rain)))) (NP (DT all) (NN day)))) (CC and) (S (NP (DT this) (是	是
6	((S (NP (PRP We)) (VP (VBD had) (S (VP (TO to) (VP (VB give) (NP (PRP it)) (PRT (RP up)) (PP (IN as) (NP (DT a) (JJ bad) (NN job)))))) (. . .)))	是	是
7	((S (NP (PRP He)) (VP (VBZ is) (NP (NP (JJ such) (DT a) (JJ bad) (NN sailor)) (SBAR (WHNP (IN that) (S (NP (PRP he)) (ADVP (RB always) (VP	是	是

Convert the Penn parsed format into the Moses required tree format.

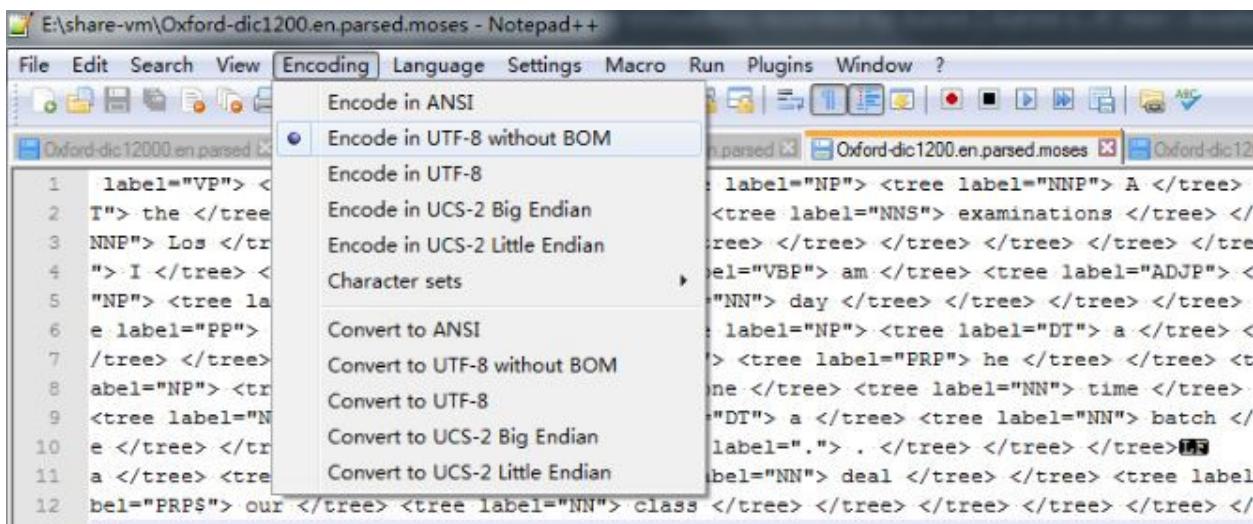
```
E:\share\vm>berkeleyparsed2mosesxml.pl < Oxford-dic1200.en.parsed > Oxford-dic1200.en.parsed.mosesxml
```

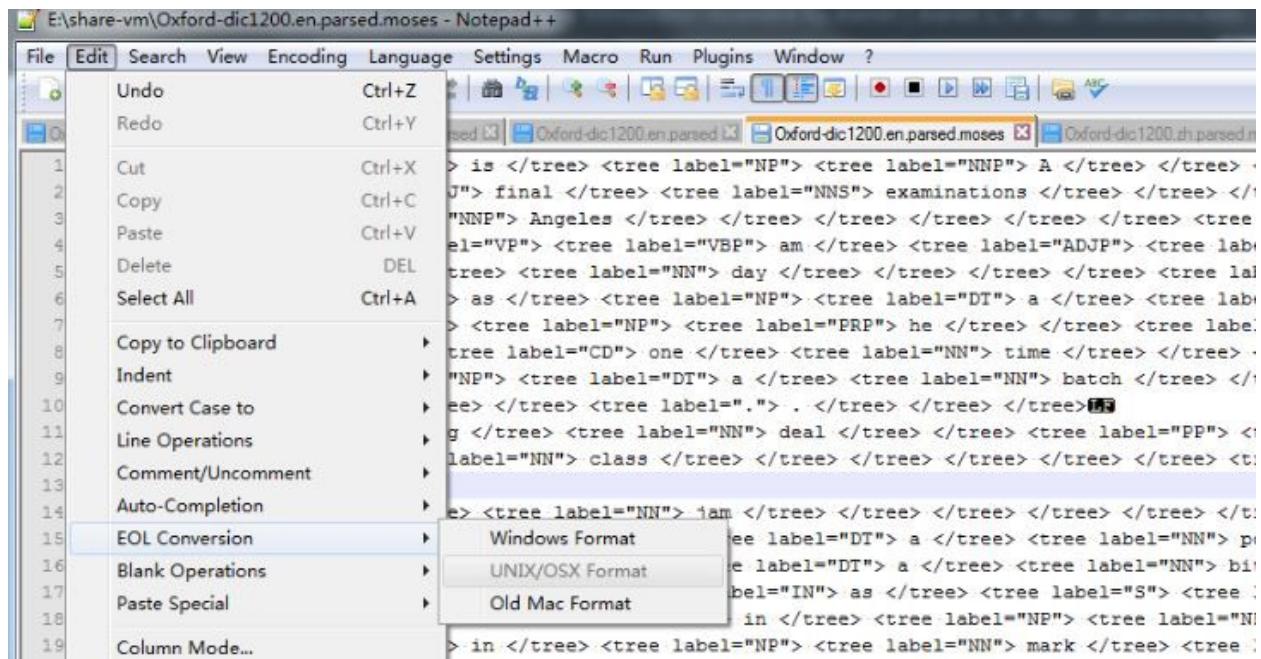
E:\share\vm>

```
E:\share-vm>berkeleyparsed2mosesxml.pl < Oxford-dic1200.zh.parsed > Oxford-dic1200.zh.parsed.mosesxml
```

After the converting, the format of the corpus becomes:

Encoding the content into TUF-8 format, edit content into UNIX format:





Training language model:

Put the converted documents “Oxford-dic1200.parsed.mosesxml.zh” and “Oxford-dic1200.parsed.mosesxml.en” into the VM directory: ~Aaron/corpus/training/

We first try to use the previously built language model “news-commentary-v8.fr-en.blm.en” from “news-commentary-v8.fr-en.tok.en”, WMT13 training corpus.

Training translation model:

Put the converted documents “Oxford-dic1200.parsed.mosesxml.zh” and “Oxford-dic1200.parsed.mosesxml.en” into the VM directory: ~Aaron/corpus/training/

Clear (empty) the directory ~Aaron/working/train/

Training command: (-f: source language; -e: target language)(without the reordering process in tree-to-tree translation training)

```
cd ~/Aaron
mkdir working
cd working
```

```

~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir ./train -corpus
~/Aaron/corpus/training/Oxford-dic1200.parsed.mosesxml -f zh -e en -alignment grow-diag-
final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm
0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir
~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.zh-en1 &

```

```

nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir ./train -corpus ~/Aaron/corpus/training/Oxford-dic1200.parsed.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.zh-en1 &

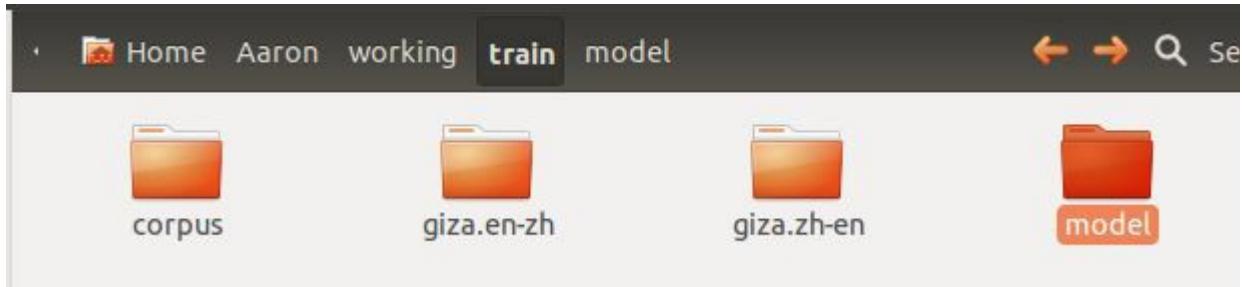
```

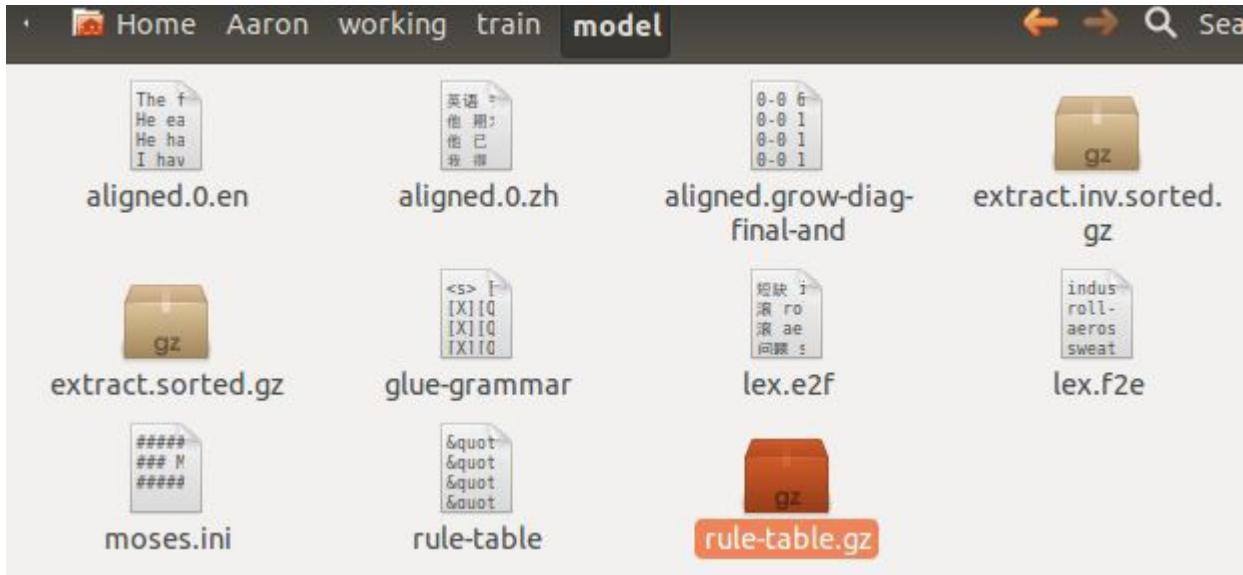
The training will be finished from 2013.11.26-17:48 to 2013.11.26-17:49. Below files will be generated:

```

mosesxml2berkeleyparsed.perl  parse-en-collins.perl  rule-table  training.out.Oxford.parse.zh-en  training.out.Oxford.parse.zh-en1
/home/nlp2ct/Aaron/Moses/mosesdecoder/scripts/../bin(score /home/nlp2ct/Aaron/working/train/model/tmp.3774/extract.0.gz /home/nlp2ct/Aaron/working/train/model/lex.f2e /home/nlp2ct/Aaron/working/train/model/tmp.3774/phrase-table.half.00000.gz --Hierarchical 2>> /dev/stderr
/home/nlp2ct/Aaron/working/train/model/tmp.3774/run.0.shmv /home/nlp2ct/Aaron/working/train/model/tmp.3774/phrase-table.half.00000.gz /home/nlp2ct/Aaron/working/train/model/rule-table.half.f2e.gzrm -rf /home/nlp2ct/Aaron/working/train/model/tmp.3774
Finished Tue Nov 26 17:49:07 2013
(6.3) creating table half /home/nlp2ct/Aaron/working/train/model/rule-table.half.e2f @ Tue Nov 26 17:49:07 CST 2013
/home/nlp2ct/Aaron/Moses/mosesdecoder/scripts/generic/score-parallel.perl 1 "sort" "/home/nlp2ct/Aaron/Moses/mosesdecoder/scripts/../bin(score /home/nlp2ct/Aaron/working/train/model/extract.inv.sorted.gz /home/nlp2ct/Aaron/working/train/model/lex.e2f /home/nlp2ct/Aaron/working/train/model/rule-table.half.e2f.gz --Inverse -Hierarchical 1
Executing: /home/nlp2ct/Aaron/Moses/mosesdecoder/scripts/generic/score-parallel.perl 1 "sort" "/home/nlp2ct/Aaron/Moses/mosesdecoder/scripts/../bin(score /home/nlp2ct/Aaron/working/train/model/extract.inv.sorted.gz /home/nlp2ct/Aaron/working/train/model/lex.e2f /home/nlp2ct/Aaron/working/train/model/rule-table.half.e2f.gz --Inverse -Hierarchical 1
Started Tue Nov 26 17:49:07 2013
ln -s /home/nlp2ct/Aaron/working/train/model/extract.inv.sorted.gz /home/nlp2ct/Aaron/working/train/model/tmp.3785/extract.0.gz
/home/nlp2ct/Aaron/Moses/mosesdecoder/scripts/../bin(score /home/nlp2ct/Aaron/working/train/model/tmp.3785/extract.0.gz /home/nlp2ct/Aaron/working/train/model/lex.e2f /home/nlp2ct/Aaron/working/train/model/tmp.3785/phrase-table.half.00000.gz --Inverse --Hierarchical 2>> /dev/stderr
/home/nlp2ct/Aaron/working/train/model/tmp.3785/run.0.shgunzip -c /home/nlp2ct/Aaron/working/train/model/tmp.3785/phrase-table.half.*.gz 2>> /dev/stderr LC_ALL=C sort -T /home/nlp2ct/Aaron/working/train/model/tmp.3785 | gzip -c > /home/nlp2ct/Aaron/working/train/model/rule-table.half.e2f.gz 2>> /dev/stderr rm -rf /home/nlp2ct/Aaron/working/train/model/tmp.3785
Finished Tue Nov 26 17:49:07 2013
(6.6) consolidating the two halves @ Tue Nov 26 17:49:07 CST 2013
Executing: /home/nlp2ct/Aaron/Moses/mosesdecoder/scripts/../bin/consolidate /home/nlp2ct/Aaron/working/train/model/rule-table.half.f2e.gz /home/nlp2ct/Aaron/working/train/model/rule-table.half.e2f.gz /dev/stdout --Hierarchical | gzip -c > /home/nlp2ct/Aaron/working/train/model/rule-table.gz
Consolidate v2.0 written by Philipp Koehn
consolidating direct and indirect rule tables
processing hierarchical rules
Executing: rm -f /home/nlp2ct/Aaron/working/train/model/rule-table.half.*
(7) learn reordering model @ Tue Nov 26 17:49:07 CST 2013
... skipping this step, reordering is not lexicalized ...
(8) learn generation model @ Tue Nov 26 17:49:07 CST 2013
no generation model requested, skipping step
(9) create moses.ini @ Tue Nov 26 17:49:07 CST 2013

```





```

train-model.perl * parse-de-berkeley.perl * mosesxml2berkeleyparsed perl * parse-en-collins perl * rule-table *
" ; [NP][NP] " ; [NP] ||| " ; [NP][NP] " ; [NP] ||| 0.42857 0.790123 1 0.790123 ||| 0-0 1-1 2-2 ||| 1.16667 0.5 0.5 |||
" ; [PU] ||| " ; [SYM] ||| 1 0.88889 0.666667 0.888889 ||| 0-0 ||| 4 6 4 ||
" ; [PU] ||| " ; [VB] ||| 1 0.88889 0.166667 0.888889 ||| 0-0 ||| 1 6 1 ||
" ; [PU] ||| will [MD] ||| 0.0666667 0.0357143 0.166667 0.111111 ||| 0-0 ||| 15 6 1 ||
" ; I [VP][VP] . [IP] ||| I will [VP][VP] . [S] ||| 0.5 0.0311261 0.5 0.0679012 ||| 0-1 1-0 2-2 3-3 ||| 0.5 0.5 0.25 ||
" ; 我 [VP][VP] . [IP] ||| I will [VP][VP] . [TOP] ||| 0.5 0.0311261 0.5 0.0679012 ||| 0-1 1-0 2-2 3-3 ||| 0.5 0.5 0.25 ||
" ; 我 [VP][VP] . [TOP] ||| I will [VP][VP] . [S] ||| 0.5 0.0311261 0.5 0.0679012 ||| 0-1 1-0 2-2 3-3 ||| 0.5 0.5 0.25 ||
" ; 我 [VP][VP] . [TOP] ||| I will [VP][VP] . [TOP] ||| 0.5 0.0311261 0.5 0.0679012 ||| 0-1 1-0 2-2 3-3 ||| 0.5 0.5 0.25 ||
" ; 玛蒂 小姐 " ; [NP] ||| &quot; Miss Mattie &quot; ; [NP] ||| 0.6 0.790123 1 0.790123 ||| 0-0 1-1 2-2 3-3 ||| 0.833333 0.5 0.5 ||
-- [PU] ||| an afterthought [NP] ||| 1 0.257634 1 0.0816327 ||| 0-0 0-1 ||| 1 1 1 ||
. [PU] ||| . [.] ||| 0.00978648 0.0149385 1 0.586207 ||| 0-0 ||| 1124 11 11 ||
1000 [QP][CD] [QP] ||| up to $ 10 [QP][CD] [NP] ||| 1 0.513514 0.5 0.88481e-06 ||| 0-0 0-3 1-4 ||| 0.333333 0.666666 0.333333 ||
1000 [QP][CD] [QP] ||| up to $ 10 [QP][CD] [QP] ||| 1 0.513514 0.5 0.88481e-06 ||| 0-0 0-3 1-4 ||| 0.333333 0.666666 0.333333 ||
1000 万 元 [QP] ||| up to $ 10 million [QP] ||| 1 0.0962838 1 2.1515e-06 ||| 0-0 0-3 1-4 2-4 ||| 0.333333 0.333333 0.333333 ||
10% [CD] ||| I [PRP] ||| 0.00934579 0.009099 1 1 ||| 0-0 ||| 107 1 1 ||
10% 的 [DNP] ||| I [PRP] ||| 0.00934579 0.00213274 1 1 ||| 0-0 ||| 107 1 1 ||
1988年 [NT] ||| 1988 [CD] ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 ||
20820 人 [NP] ||| numbered 20820 [VP] ||| 1 0.375 1 0.127358 ||| 0-0 0-1 1-0 ||| 1 1 1 ||
980 [CD] ||| 703,980 [CD] ||| 1 1 1 1 ||| 0-0 ||| 1 1 1 ||
: [PU] ||| as [RB] ||| 0.2 0.0163934 1 0.5 ||| 0-0 ||| 5 1 1 ||
; [PU] ||| visit [NN] ||| 1 0.333333 1 1 ||| 0-0 ||| 1 1 1 ||
[ADVP][TO] 克服 [IP] ||| [ADVP][TO] crack [SBAR] ||| 0.181818 0.5 0.333333 1 ||| 0-0 1-1 ||| 0.785714 0.428571 0.142857 ||
[ADVP][TO] 克服 [IP] ||| [ADVP][TO] crack [S] ||| 0.181818 0.5 0.333333 1 ||| 0-0 1-1 ||| 0.785714 0.428571 0.142857 ||
[ADVP][TO] 克服 [IP] ||| [ADVP][TO] crack [VP] ||| 0.181818 0.5 0.333333 1 ||| 0-0 1-1 ||| 0.785714 0.428571 0.142857 ||
[ADVP][TO] 克服 [VP] ||| [ADVP][TO] crack [SBAR] ||| 0.181818 0.5 0.333333 1 ||| 0-0 1-1 ||| 0.785714 0.428571 0.142857 ||

```

Refer below command is from Tianliang:

```

Kickstarted 12:40 28-Mar-2012
[ma96572@pearl ~]$ nohup nice /home/ma96572/research/moses/scripts/training/train-model.perl -cores 5 --parallel -root-dir /home/ma96572/Michelle/Syntax-based2/ -external-bin-dir /home/ma96572/research/giza-pp/bin -corpus /home/ma96572/Michelle/Corpus/train5 -f en -e zh -alignment grow-diag-final-and -first-step 1 -last-step 9 -hierarchical -glue-grammar -lm 0:5:/home/ma96572/Michelle/LM/train5.bl m.zh:0 >& /home/ma96572/Michelle/trace/out3 &

```

```

"/home/ma96572/research/moses/scripts/training/train-model.perl -cores 5 --parallel -root-dir
/home/ma96572/Michelle/Syntax-based2/ -external-bin-dir /home/ma96572/research/giza-pp/bin
-corpus /home/ma96572/Michelle/Corpus/train5 -f en -e zh -alignment grow-diag-final-and -first-step 1

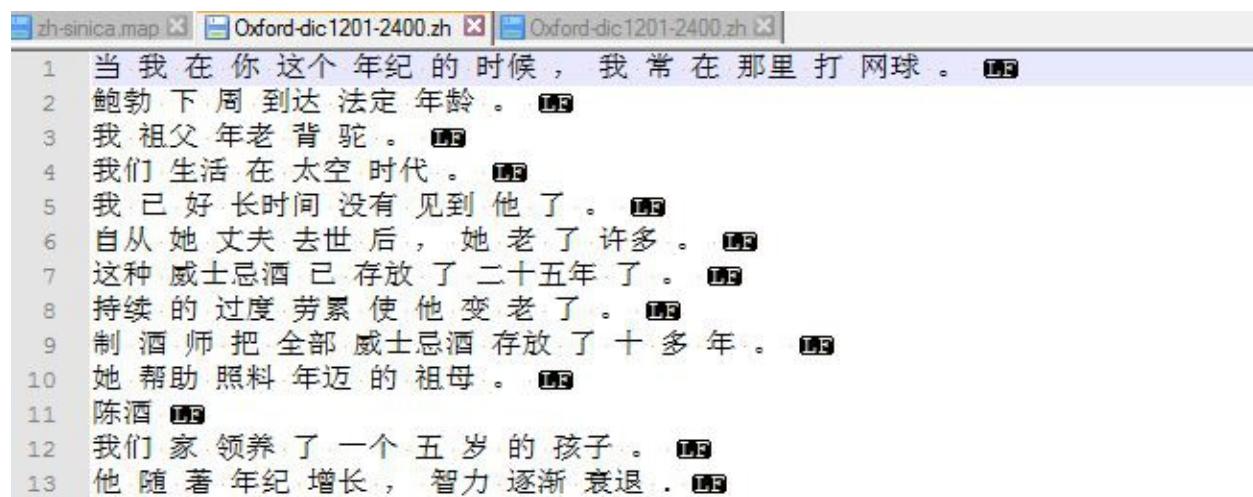
```

```
-last-step 9 -hierarchical -glue-grammar -lm 0:5:/home/ma96572/Michelle/LM/train5.blm.zh:0 >& /home/ma96572/Michelle/trace/out3"
```

Testing:

We select the 1201th to 2400th Chinese sentences from the Oxford-dic120000.segmented EN-ZH bilingual corpus for testing.

Input testing document:



```
zh-sinica.map Oxford-dic1201-2400.zh Oxford-dic1201-2400.zh
1 当 我 在 你 这 个 年 纪 的 时 候 , 我 常 在 那 里 打 网 球 。 LF
2 鲍 勃 下 周 到 达 法 定 年 龄 。 LF
3 我 祖 父 年 老 背 驼 。 LF
4 我 们 生 活 在 太 空 时 代 。 LF
5 我 已 好 长 时 间 没 有 见 到 他 了 。 LF
6 自 从 她 丈 夫 去 世 后 , 她 老 了 许 多 。 LF
7 这 种 威 士 忌 酒 已 存 放 了 二 十 五 年 了 。 LF
8 持 续 的 过 度 劳 累 使 他 变 老 了 。 LF
9 制 酒 师 把 全 部 威 士 忌 酒 存 放 了 十 多 年 。 LF
10 她 帮 助 照 料 年 迈 的 祖 母 。 LF
11 陈 酒 LF
12 我 们 家 领 养 了 一 个 五 岁 的 孩 子 。 LF
13 他 随 著 年 纪 增 长 , 智 力 逐 渐 衰 退 . LF
```

Using the testing command:

```
nlp2ct@nlp2ct-VirtualBox:~$ cd Aaron
nlp2ct@nlp2ct-VirtualBox:~/Aaron$ cd working/
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_c
hart -f train/model/moses.ini < ~/Aaron/corpus/test/tree-to-tree/Oxford-dic1201-
2400.zh > Oxford-dic1201-2400.zh-en.ttt.out
```

The translation result:

when I in your The 年纪 of 时候 , I often had teases 网球 .
 鲍勃 'll 到达 法定 年龄 .
 I 祖父 年老 himself 驼 .
 our life in 太空 时代 .
 I .Brown good 长时间 no 见到 his abutting .
 自从 adornments 丈夫 was after , adornments he abutting Many .
 这种 威士忌酒 .Brown 存放 abutting 二十五年 abutting .
 持续 of One 劳累 the his 变 he abutting .
 制 a 师 to access 威士忌酒 存放 abutting 十多 years .
 her help 照料 年迈 of 祖母 .
 陈酒
 a 领养 abutting a 五岁 of development .
 his 随著 年纪 增长 , 智力 逐渐 衰退 .
 I ask 托尼 为什么 上班 迟到 . he let me his 碰到 abutting 倒霉 things , his mother The

Here, we change the testing format. Before the testing, we parse the input chinese sentences using the traindGrammarCTB-7, the input testing document:

1 <tree label="TOP"> <trees label="IP"> <trees label="DPL"> <trees label="P"> 当 </tree> <trees label="PP"> <trees label="P"> 在 </tree> <trees label="PP"> <trees label="P"> 我 </tree> <trees label="VP"> <trees label="DP"> <trees label="DT"> </tree> <trees label="CLP"> <trees label="M"> 鲍勃 </tree> <trees label="NP"> <trees label="NNR"> 爷爷 </tree> <trees label="VP"> <trees label="VP"> <trees label="VA"> 年老 </tree> <trees label="TOP"> <trees label="IP"> <trees label="NP"> <trees label="NN"> 祖父 </tree> <trees label="VP"> <trees label="VP"> <trees label="VA"> 生活 </tree> <trees label="PP"> <trees label="P"> 在 </tree> <trees label="TOP"> <trees label="IP"> <trees label="NP"> <trees label="NP"> 我们 </tree> <trees label="VP"> <trees label="VV"> 生活 </tree> <trees label="PP"> <trees label="P"> 在 </tree> <trees label="TOP"> <trees label="IP"> <trees label="NP"> <trees label="NP"> 我 </tree> <trees label="VP"> <trees label="ADVP"> <trees label="AD"> 已 </tree> <trees label="TOP"> <trees label="IP"> <trees label="PP"> <trees label="P"> 自从 </tree> <trees label="TGP"> <trees label="IP"> <trees label="NP"> <trees label="NN"> 她 </tree> <trees label="NP"> <trees label="NN"> 丈夫 .0 <trees label="TOP"> <trees label="IP"> <trees label="NP"> <trees label="NP"> <trees label="NN"> 这种 </tree> <trees label="VP"> <trees label="NP"> <trees label="ADVP"> <trees label="TOP"> <trees label="IP"> <trees label="NP"> <trees label="NP"> <trees label="NN"> 威士忌酒 </tree> <trees label="VP"> <trees label="NP"> <trees label="NN"> 酒 </tree> <trees label="TOP"> <trees label="IP"> <trees label="NP"> <trees label="NP"> <trees label="NN"> 持续 </tree> <trees label="VP"> <trees label="NP"> <trees label="NN"> 酒 </tree> <trees label="TOP"> <trees label="IP"> <trees label="NP"> <trees label="NP"> <trees label="NN"> 帮助 </tree> <trees label="VP"> <trees label="NP"> <trees label="VV"> .1 <trees label="TOP"> <trees label="FRAG"> <trees label="NR"> 陈酒 </tree> </tree> </tree>

The testing command:

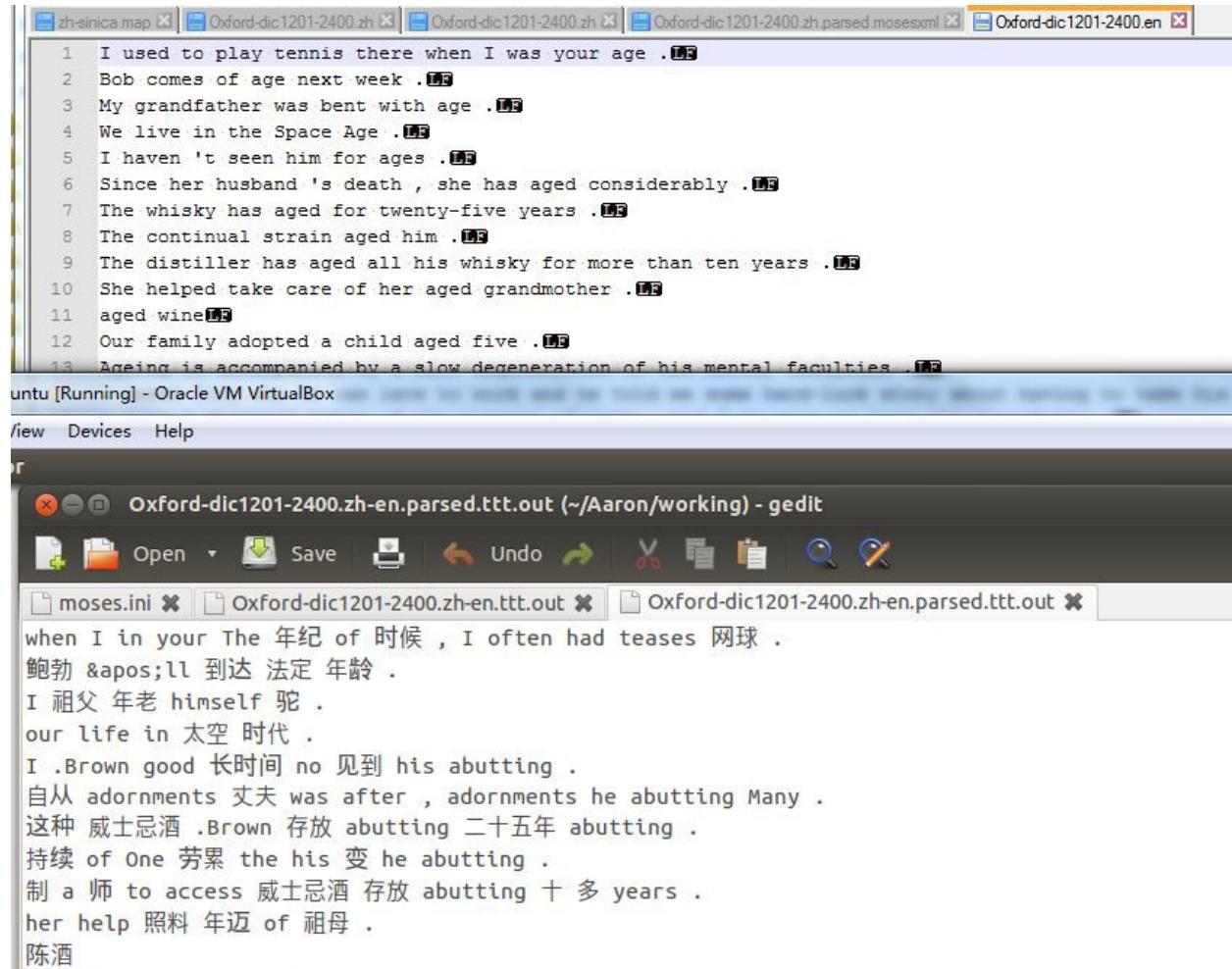
```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_chart -f train/model/moses.ini < ~/Aaron/corpus/test/tree-to-tree/Oxford-dic1201-2400.zh.parsed.mosesxml > Oxford-dic1201-2400.zh-en.parsed.ttt.out
```

The testing translation result:

when I in your The 年纪 of 时候 , I often had teases 网球 .
 鲍勃 'll 到达 法定 年龄 .
 I 祖父 年老 himself 驼 .
 our life in 太空 时代 .
 I .Brown good 长时间 no 见到 his abutting .
 自从 adornments 丈夫 was after , adornments he abutting Many .
 这种 威士忌酒 .Brown 存放 abutting 二十五年 abutting .
 持续 of One 劳累 the his 变 he abutting .
 制 a 师 to access 威士忌酒 存放 abutting 十多 years .
 her help 照料 年迈 of 祖母 .
 陈酒
 a 领养 abutting a 五岁 of development .
 his 随著 年纪 增长 , 智力 逐渐 衰退 .
 I ask 托尼 为什么 上班 迟到 . he let me his 碰到 abutting 倒霉 things , his mother The 宝贝 of 长卷毛
 a gave microwave go to 兽医 .

The translation result using the parsed input sentence is the same with previous result (without parsing the input sentence).

The comparison of automatically translated sentences and references:



The terminal window displays 13 numbered English sentences, each followed by a line separator (LF). The sentences are:

- 1 I used to play tennis there when I was your age .LF
- 2 Bob comes of age next week .LF
- 3 My grandfather was bent with age .LF
- 4 We live in the Space Age .LF
- 5 I haven 't seen him for ages .LF
- 6 Since her husband 's death , she has aged considerably .LF
- 7 The whisky has aged for twenty-five years .LF
- 8 The continual strain aged him .LF
- 9 The distiller has aged all his whisky for more than ten years .LF
- 10 She helped take care of her aged grandmother .LF
- 11 aged wineLF
- 12 Our family adopted a child aged five .LF
- 13 Ageing is accompanied by a slow degeneration of his mental faculties .LF

Below the terminal window is a screenshot of a text editor (gedit) showing the same 13 sentences in Chinese. The editor interface includes a toolbar with file operations like Open, Save, Undo, and Redo, and a menu bar with View, Devices, Help.

when I in your The 年纪 of 时候 , I often had teases 网球 .
鲍勃 'll 到达 法定 年龄 .
I 祖父 年老 himself 驼 .
our life in 太空 时代 .
I .Brown good 长时间 no 见到 his abutting .
自从 adornments 丈夫 was after , adornments he abutting Many .
这种 威士忌酒 .Brown 存放 abutting 二十五年 abutting .
持续 of One 劳累 the his 变 he abutting .
制 a 师 to access 威士忌酒 存放 abutting 十 多 years .
her help 照料 年迈 of 祖母 .
陈酒 .

Preparing larger corpus-2

Using the Oxford-dictionary.zh 12000 tokenized simplified Chinese sentences, the GrammarTrained on CTB-7, BerkeleyParser-1.7.jar, the parsing information: 10 minutes.

Using the TrainedGramEng-WSJ, Berkeley-Parser-1.7.jar, Oxford-dictionary.zh 12000 tokenized English sentences, the parsing: 11 minutes.

10000 sentences (1st-10000th) for training, 1000 sentences (10001th-11000th) for developing, 1000 (11001th-12000th) sentences for testing. We skip the developing stage this time.

Put the files “Oxford-dic10000.parsed.mosesxml.en” and “Oxford-dic10000.parsed.mosesxml.zh” in the directory “~/Aaron/corpus/training/train10000Oxf”.

Training translation model-2:

```
cd ~/Aaron  
mkdir working  
cd working  
~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000 -corpus  
~/Aaron/corpus/training/train10000Oxf/Oxford-dic10000.parsed.mosesxml -f zh -e en  
-alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm  
0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir  
~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.zh-en10000 &
```

```
nlp2ct@nlp2ct-VirtualBox:~$ cd Aaron  
nlp2ct@nlp2ct-VirtualBox:~/Aaron$ cd working  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000 -corpus ~/Aaron/corpus/training/train10000Oxf/Oxford-dic10000.parsed.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.zh-en10000 &
```

It runs as below:

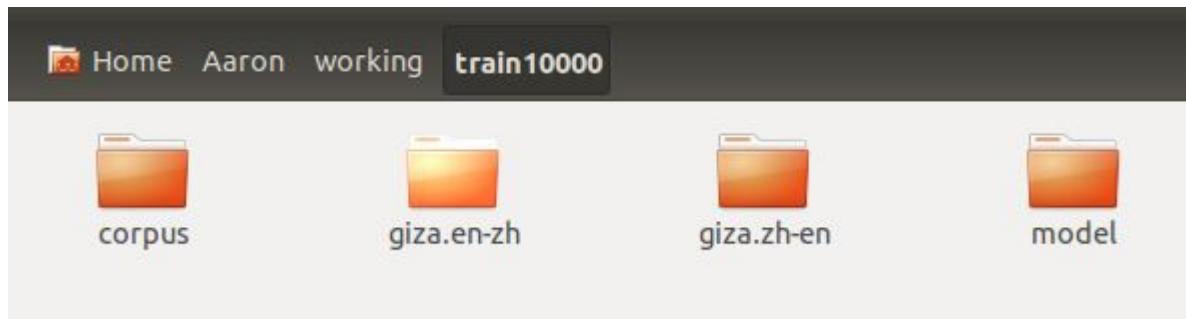
```

nlp2ct@nlp2ct-VirtualBox: ~/Aaron/working
top - 13:42:28 up 2:54, 2 users, load average: 0.36, 0.11, 0.07
Tasks: 153 total, 2 running, 151 sleeping, 0 stopped, 0 zombie
Cpu(s): 52.6%us, 3.1%sy, 0.0%ni, 44.3%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 6786596k total, 1138116k used, 5648480k free, 47352k buffers
Swap: 7336956k total, 0k used, 7336956k free, 450752k cached

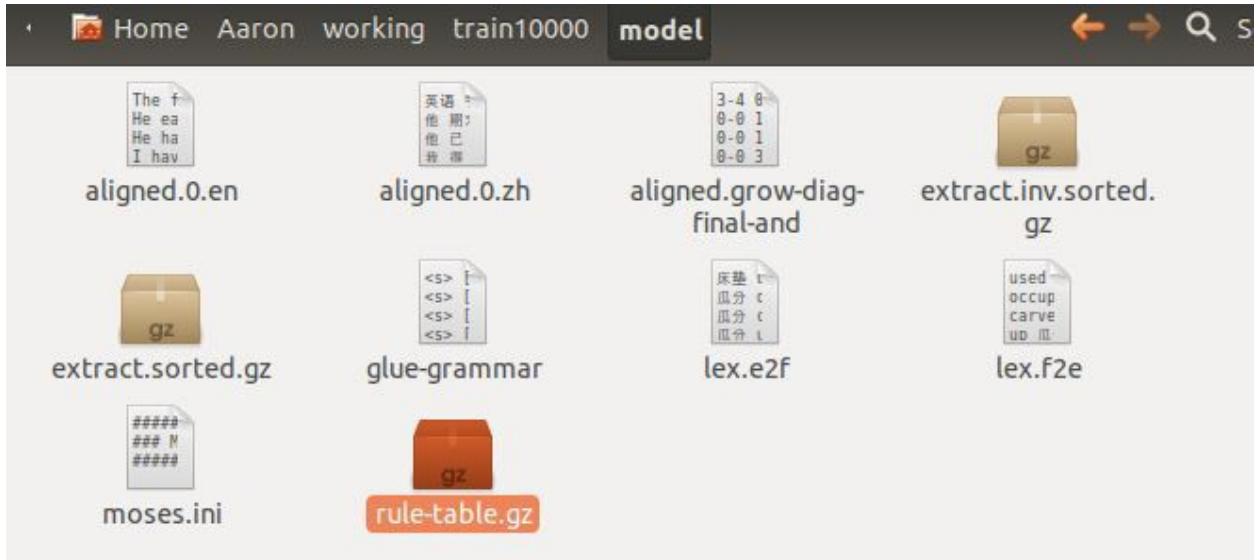
PID USER      PR NI  VIRT   RES   SHR S %CPU %MEM     TIME+ COMMAND
2370 nlp2ct    20  0 22356  10m 1396 R  49  0.2  0:01.47 mkcls
1226 root      20  0 315m 138m 14m S   4  2.1  1:20.15 Xorg
1656 nlp2ct    20  0 1484m 92m 36m S   2  1.4  1:30.53 compiz
1679 nlp2ct    9 -11 354m 6228 3908 S   2  0.1  0:00.47 pulseaudio
1690 nlp2ct    20  0 1227m 39m 20m S   1  0.6  0:11.99 nautilus
1617 nlp2ct    20  0 109m 1568 1012 S   0  0.0  0:29.25 VBoxClient
1783 nlp2ct    20  0 383m 9.8m 7336 S   0  0.1  0:00.66 bamfdaemon
1792 nlp2ct    20  0 312m 10m 8408 S   0  0.2  0:00.96 gtk-window-deco
1961 nlp2ct    20  0 407m 5952 4568 S   0  0.1  0:00.38 zeitgeist-datab
2052 nlp2ct    20  0 307m 9m 7600 S   0  0.2  0:01.50 gnome-screensav
2202 nlp2ct    20  0 511m 17m 11m S   0  0.3  0:01.29 gnome-terminal
2329 nlp2ct    20  0 630m 25m 14m S   0  0.4  0:01.29 gedit
  1 root       20  0 24468 2412 1352 S   0  0.0  0:01.05 init
  2 root       20  0      0    0 S   0  0.0  0:00.00 kthreadd
  3 root       20  0      0    0 S   0  0.0  0:01.16 ksoftirqd/0
  5 root       20  0      0    0 S   0  0.0  0:00.80 kworker/u:0
  6 root       RT  0      0    0 S   0  0.0  0:00.01 migration/0

```

The training generates the files:



The rule-table.gz is 1.2MB large.



```

record-moses-commands02.txt ✘ rule-table ✘
! [PU] ||| ! [.] ||| 0.0375 0.198529 1 0.794118 ||| 0-0 ||| 80 3 3 |||
&quot; A &quot; [QP] ||| letter [NN] ||| 0.0833333 7.76749e-09 0.5 0.5 ||| 1-0 ||| 12 2 1 |||
&quot; A &quot; [QP] ||| the letter [NP] ||| 0.333333 7.76749e-09 0.5 0.0608267 ||| 1-1 ||| 3 2 1 |||
&quot; [IP][TO] 即付 &quot; [IP] ||| [IP][TO] bearer on demand &quot; [S] ||| 1 1.43341e-11 0.5 2.40744e-11
&quot; [IP][TO] 即付 &quot; [IP] ||| [IP][TO] bearer on demand &quot; [VP] ||| 1 1.43341e-11 0.5 2.40744e-11
&quot; [PU] ||| &quot; [NN] ||| 1 0.387755 0.0833333 0.44186 ||| 0-0 ||| 1 12 1 |||
&quot; [PU] ||| &quot; [SYM] ||| 0.470588 0.387755 0.666667 0.44186 ||| 0-0 ||| 17 12 8 |||
&quot; [PU] ||| &quot; [VB] ||| 1 0.387755 0.0833333 0.44186 ||| 0-0 ||| 1 12 1 |||
&quot; [PU] ||| bearer on demand &quot; [VP] ||| 1 0.387755 0.0833333 2.40744e-11 ||| 0-3 ||| 1 12 1 |||
&quot; [PU] ||| notice [NN] ||| 1 0.166667 0.0833333 0.0232558 ||| 0-0 ||| 1 12 1 |||
&quot; [VP][TO] 即付 &quot; [IP] ||| [VP][TO] bearer on demand &quot; [S] ||| 1 1.43341e-11 0.5 2.40744e-11
&quot; [VP][TO] 即付 &quot; [IP] ||| [VP][TO] bearer on demand &quot; [VP] ||| 1 1.43341e-11 0.5 2.40744e-11
&quot; p &quot; [NP] ||| to [TO] ||| 0.00162075 1.12243e-10 1 1 ||| 1-0 ||| 617 1 1 |||
&quot; 快 [IP] ||| Fast [IN] ||| 1 0.25 1 0.0229915 ||| 0-0 1-0 ||| 1 1 1 |||
&quot; 我 [VP][NP] 。 [IP] ||| I will &quot; is [VP][NP] . [S] ||| 0.5 0.320969 0.5 1.40163e-05 ||| 0-2 1-0 2-
&quot; 我 [VP][NP] 。 [IP] ||| I will &quot; is [VP][NP] . [TOP] ||| 0.5 0.320969 0.5 1.40163e-05 ||| 0-2 1-0
&quot; 我 [VP][NP] 。 [TOP] ||| I will &quot; is [VP][NP] . [TOP] ||| 0.5 0.320969 0.5 1.40163e-05 ||| 0-2 1-0

```

Testing-2

We select the 11001 to 12000 sentences for testing.

Input testing document:

tree label="TOP" <tree label="IP" <tree label="NP" <tree label="NR" 罗马 /><tree label="VP" <tree label="VC"> 是 /><tree label="NP" <tree label="TOP" <tree label="IP" <tree label="NP" <tree label="PN" 她 /></tree><tree label="VP" <tree label="ADVP"><tree label="AD"> 已 /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="NN" 十字架 /></tree><tree label="VP" <tree label="VC"> 是 /><tree><tree label="NP" <tree label="TOP" <tree label="IP" <tree label="NP" <tree label="PN" 他 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 他 /><tree><tree label="VP" <tree label="NP" <tree label="NN" 去 /></tree><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="CNP"><tree label="NP" <tree label="NN" 十字架 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 主席 /><tree label="VP" <tree label="ADVP"><tree label="AD"> 已 /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="VP" <tree label="VW"> 苏赤 /><tree><tree label="NP" <tree label="NN" 圣诞 /></tree><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="NN" 人们 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 过去 /></tree><tree label="NP" <tree label="NN" 在 /></tree><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" NT"> 在过去 /><tree><tree label="DEG"> 的 /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" NT"> 今年 /></tree><tree label="DEG"> 的 /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="DEP"><tree label="DT"> 那 /></tree><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="NN" 圣诞节 /></tree><tree label="NP" <tree label="NN" 火车 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 圣诞节 /></tree><tree label="NP" <tree label="NN" 前夕 /></tree><tree label="NP" <tree label="NN" 节 /><tree><tree label="NP" <tree label="NN" 期 /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" 我们 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 去 /></tree><tree label="NP" <tree label="NN" 圣诞 /><tree><tree label="NP" <tree label="NN" 他们 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 着力 /><tree><tree label="NP" <tree label="NN" 圣诞 /><tree><tree label="NP" <tree label="NN" 我们 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 买 /><tree><tree label="NP" <tree label="NN" 有人 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 按 /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" 圣诞树 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 孩子 /></tree><tree label="NP" <tree label="NN" 们 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 着 /><tree><tree label="NP" <tree label="NN" 伸长 /><tree><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" 烛光 /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" 他们 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 用 /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="PN" 她 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 克里斯托弗·哥伦布 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 是 /><tree><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="PN" 我 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 伸长 /><tree><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" 基因 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 座 /><tree><tree label="NP" <tree label="NN" 尽管 /><tree><tree label="NP" <tree label="NN" 他 /></tree><tree label="VP" <tree label="NP" <tree label="NN" 他 /><tree><tree label="NP" <tree label="NN" 胖乎乎 /><tree><tree label="NP" <tree label="NN" 胖乎乎 /><tree label="DEG"> 的 /><tree><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" 那 /></tree><tree label="VP" <tree label="ADVP"><tree label="AD"> AD /><tree /><tree label="TOP" <tree label="IP" <tree label="NP" <tree label="DNP"><tree label="NP" <tree label="NN" 抚养 /></tree><tree label="NP" <tree label="NN" 婴儿 /></tree><tree label="NP" <tree label="NN" 下巴 /></tree><tree label="DEG"> 的 />

Using the testing command:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_c  
hart -f train10000/model/moses.ini < ~/Aaron/corpus/test/tree-to-tree/Oxford-dic  
11001-12000.zh.parsed.mosesxml > Oxford-dic11001-12000.zh.parsed.ttt-out
```

The translation result:

```
          1   0   0
          5   0
          1
BEST TRANSLATION: 170 Q -> Q </s> :0-0 : c=-0.460 core=(0.000,-1.000,1.000,0.00
0,0.000,0.000,0.000,0.000) [0..18] 168 [total=-322.267] core=(-300.000,-1
8.000,34.000,-20.515,-30.775,-8.093,-17.636,15.998,-95.322)
Translation took 0.030 seconds
End. : [19.000] seconds
Name:moses_chart      VmPeak:236388 kB      VmRSS:98156 kB  RSSMax:99492 kB
ser:1.976      sys:8.405      CPU:10.381      real:19.320
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

The comparison of automatical translation and reference sentences:

Roman is that the world the the great city .
her has皈依基督教 .
十字架 is 基督教 of 象征 .
he to foreign bows 异教徒 宣传 基督教 .
he recently 改 letter 基督教 the .
基督教 义 welding of 三位一体 指 of is 圣父 , 圣子 and 圣灵 .
十字架 is 基督教 of as 徵 .
the has to 克里斯蒂娜 make his 继任 people .
恭祝 圣诞 , and 贺新禧 ! .
people in Christmas 互 staff 贺卡 and present .
Grain of robbers Christmas holiday been Gatwick quirk at peace .
this year of Christmas is 星期一 .
the country 's people across Christmas you ?
Christmas train 停驶 .
Christmas billows eat 火鸡 is in England of traditional .
Christmas 前夕 , Mr. Smith was glad , because he received friends of many by .

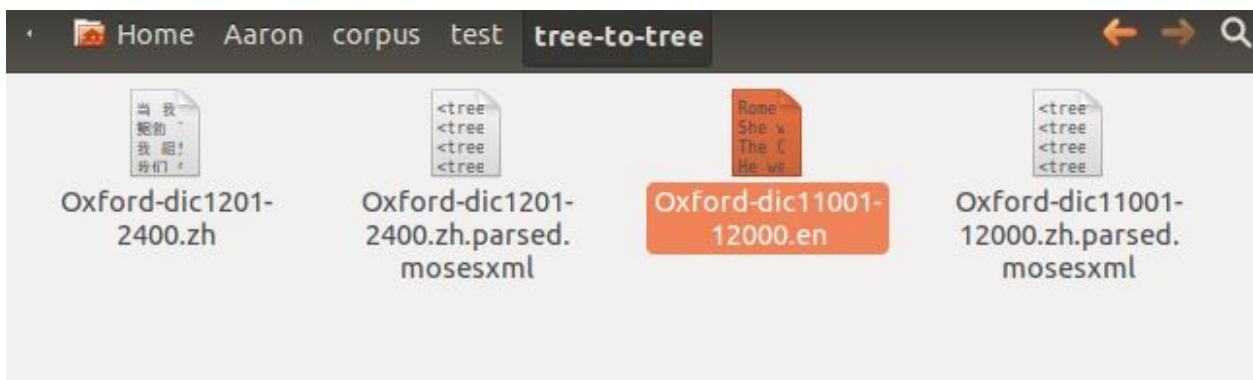
E:\Berkeley_Parser\corpus\120000-corpus\split\Oxford-dic120000.en - Notepad++

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?

11001 Rome was the greatest city in all Christendom .LF
11002 She was converted to Christianity .LF
11003 The Cross is the symbol of Christianity .LF
11004 He went abroad to preach Christianity to the heathen .LF
11005 He is a recent convert to Christianity .LF
11006 The trinity in Christianity refers to the union of the Father , the Son and the Holy Spirit .LF
11007 The cross is symbolic of Christianity .LF
11008 The chairman has designated Christina as his successor .LF
11009 Merry Christmas and a Happy New Year ! LF
11010 People send each other cards and presents at Christmas .LF
11011 The past few Christmases were very quiet .LF
11012 Christmas Day falls on a Monday .LF
11013 Do they observe Christmas Day in that country ?LF
11014 The trains don 't run on Christmas Day .LF
11015 It 's traditional in England to eat turkey on Christmas Day .LF
11016 Mr Smith was very happy on Christmas Eve because he received a batch of letters from his friends .LF

Calculate the testing score using BLEU:

Put the reference document:



Using the following command, we calculate the BLEU score of the 1000 translated sentences:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/gen  
eric/multi-bleu.perl -lc ~/Aaron/corpus/test/tree-to-tree/Oxford-dic11001-12000.  
en < Oxford-dic11001-12000.zh.parsed.ttt-out
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/gen  
eric/multi-bleu.perl -lc ~/Aaron/corpus/test/tree-to-tree/Oxford-dic11001-12000.  
en < Oxford-dic11001-12000.zh.parsed.ttt-out  
BLEU = 4.07, 36.2/6.7/1.7/0.6 (BP=1.000, ratio=1.005, hyp_len=9951, ref_len=9905  
)  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Preparing larger corpus-2-Uni-phrase-tags

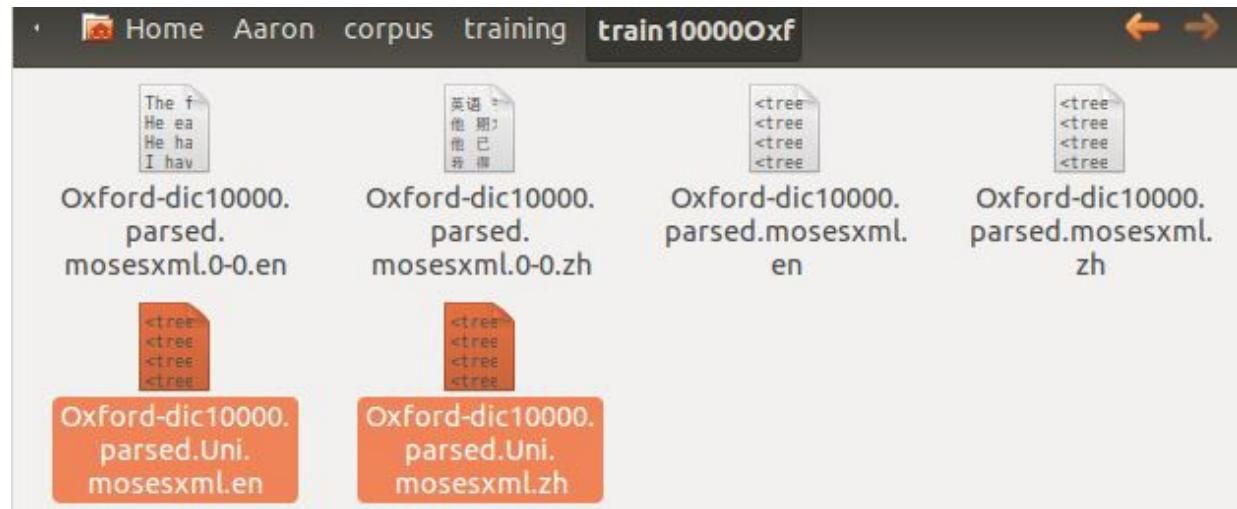
Using the Oxford-dictionary.zh 12000 tokened simplified Chinese sentences, the GrammarTrained on CTB-7, BerkeleyParser-1.7.jar, the parsing information: 10 minutes.

Using the TrainedGramEng-WSJ, Berkeley-Parser-1.7.jar, Oxford-dictionary.zh 12000 tokened English sentences, the parsing: 11 minutes.

Replace the phrase tagset in parsed 12000 bilingual EN-ZH sentences using the universal phrase tagset.

10000 sentences (1st-10000th) for training, 1000 sentences (10001th-11000th) for developing, 1000 (11001th-12000th) sentences for testing. We skip the developing stage this time.

Put the files “Oxford-dic10000.parsed.Uni.mosesxml.zh” and “Oxford-dic10000.parsed.Uni.mosesxml.en” in the directory “~/Aaron/corpus/training/train10000Oxf”.



```

[Oxford-dic10000.en.parsed.Uni.mosesxml] [Oxford-dic10000.zh.parsed.Uni.mosesxml]
1 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UADV"> <tree label="UNP"> <tree label="UNP"> <tree label="NN"> 英语 </tree> <tree label="TOP">
2 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 他 </tree> </tree> <tree label="UVB"> <tree label="UNP"> <tree label="UNP"> <tree label="AD"> 已 </t>
3 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 他 </tree> </tree> <tree label="UVB"> <tree label="UADV"> <tree label="AD"> 已 </t>
4 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 我 </tree> </tree> <tree label="UVB"> <tree label="UVB"> 得 </t>
5 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 他 </tree> </tree> <tree label="UVB"> <tree label="UVB"> <tree label="UADV"> <tree label="UNP"> <tree label="UNP"> 我 </t>
6 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 我们 </tree> </tree> <tree label="UVB"> <tree label="UADV"> <tree label="AD"> 只 </t>
7 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 他 </tree> </tree> <tree label="UVB"> <tree label="VC"> 是 </tree>
8 <tree label="TOP"> <tree label="US"> <tree label="UVB"> <tree label="UNP"> <tree label="UVB"> <tree label="UADV"> <tree label="UVB"> <tree label="UNP"> <tree label="UNP"> 我 </t>
9 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> <tree label="UADV"> <tree label="UNP"> <tree label="UNP"> <tree label="UNP"> 我 </t>
10 <tree label="TOP"> <tree label="US"> <tree label="UVB"> <tree label="UNP"> <tree label="US"> <tree label="UNP"> <tree label="NN"> 会计 </t>
11 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> <tree label="UNP"> 这 </tree> </tree> <tree label="UVB"> <tree label="UADV"> <tree label="UNP">
12 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="P"> 在 </tree> <tree label="UNP"> <tree label="UNP"> <tree label="UNP">
13 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UADV"> <tree label="UNP"> <tree label="UNP"> 你 </tree> </tree> <tree label="DEG"> 的 </t>
14 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 我 </tree> <tree label="NN"> 手头 </tree> </tree> <tree label="UNP">
15 <tree label="TOP"> <tree label="US"> <tree label="UVB"> <tree label="VV"> 要 </tree> <tree label="UVB"> <tree label="VV"> 想 </tree> <tree label="UNP">
16 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 他们 </tree> </tree> <tree label="UVB"> <tree label="UADV"> <tree label=""> 断言 </t>
17 <tree label="TOP"> <tree label="US"> <tree label="UNP"> <tree label="UNP"> 他 </tree> </tree> <tree label="UVB"> <tree label="VV"> 打开 </t>

```

Training translation model-2-Uni-phrase-tags

Using the following command to train the translation model:

```

nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000Uni -corpus ~/Aaron/corpus/training/train10000xf/Oxford-dic10000.parsed.Uni.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:SHOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.Uni.zh-en10000 &
[1] 2628
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ top
top - 15:59:01 up 5:11, 2 users, load average: 0.45, 0.13, 0.08
Tasks: 153 total, 2 running, 151 sleeping, 0 stopped, 0 zombie
Cpu(s): 51.3%us, 0.2%sy, 0.0%ni, 48.6%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 6786596k total, 1309460k used, 5477136k free, 49668k buffers
Swap: 7336956k total, 0k used, 7336956k free, 559804k cached

```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
2660	nlp2ct	20	0	49208	36m	1968	R	100	0.5	0:08.94	GIZA++
1226	root	20	0	319m	141m	14m	S	1	2.1	2:23.45	Xorg
1617	nlp2ct	20	0	109m	1568	1012	S	1	0.0	0:51.01	VBoxClient
2202	nlp2ct	20	0	528m	23m	15m	S	1	0.4	0:19.54	gnome-terminal
1655	clash	20	0	1159m	92m	26m	S	1	1.1	2:29.57	curl

It finished as:

```

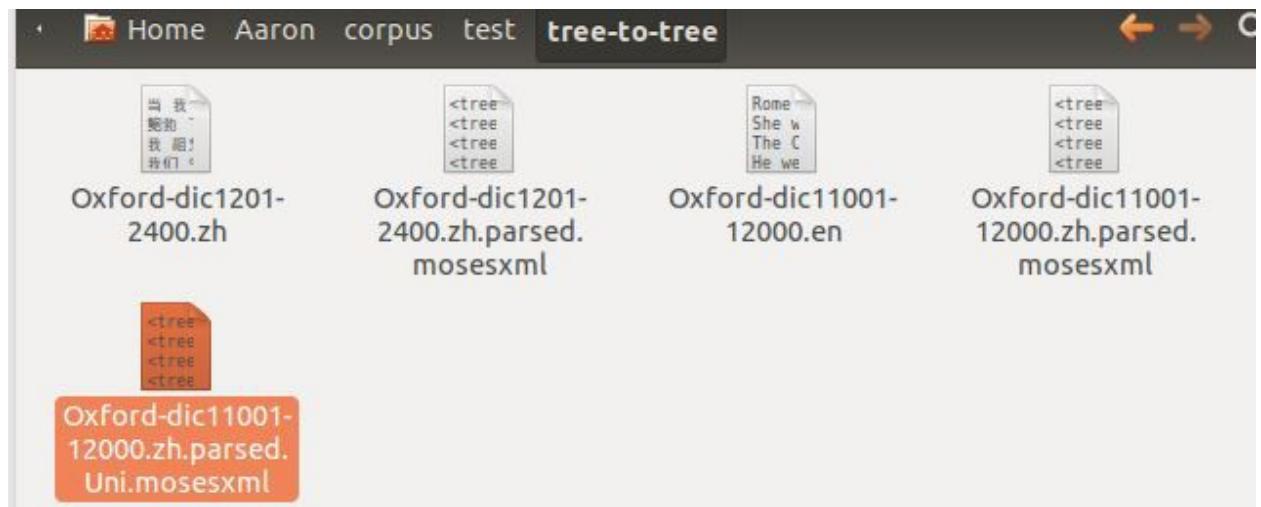
15 root      0 -20      0    0    0 S    0  0.0  0:00.00 netns
[1]+ Done                  ~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000Uni -corpus ~/Aaron/corpus/training/train10000xf/Oxford-dic10000.parsed.Uni.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:SHOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >&training.out.Oxford.parse.Uni.zh-en10000
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 

```



Testing-2-Uni-phrase-tags

Put the testing document:



Using the following command for testing:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_c  
hart -f train10000Uni/model/moses.ini < ~/Aaron/corpus/test/tree-to-tree/Oxford-  
dic11001-12000.zh.parsed.Uni.mosesxml > Oxford-dic11001-12000.zh.parsed.Uni.ttt-  
out
```

It finished as:

```
 15 root      0 -20     0   0   0 S    0 0.0   0:00.00 netns
[1]+ Done                  ~/Aaron/Moses/mosesdecoder/scripts/training/train-
model.perl -root-dir train10000Uni -corpus ~/Aaron/corpus/training/train10000xf
/Oxford-dic10000.parsed.Uni.mosesxml -f zh -e en -alignment grow-diag-final-and
-hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/
lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecode
r/tools &>training.out.Oxford.parse.Uni.zh-en10000
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ █
```

```
Oxford-dic11001-12000.zh.parsed.ttt-out ✘ Oxford-dic11001-120...h.parsed.Uni.ttt-out ✘
Roman is that the world the the great city .
her has 基督教 .
十字架 is 基督教 of 象征 .
he to foreign bows 异教徒 宣传 基督教 .
he recently 改 letter 基督教 the .
基督教 义 welding of 三位一体 指 of is 圣父 , 圣子 and 圣灵 .
十字架 is 基督教 of as 徵 .
the has to 克里斯蒂娜 make his 继任 people .
恭祝 圣诞 , and 贺新禧 ! .
people in Christmas 互 staff 贺卡 and present .
Grain of robbers Christmas holiday been Gatwick quirk at peace .
this year of Christmas is 星期一 .
that a country of people across Christmas you ?
Christmas train 停驶 .
Christmas billows eat 火鸡 is in England of traditional .
Christmas 前夕 , Mr. Smith was glad , because he received friends of many by .
Christmas 节期 圣诞 节假日 from 十二月二十四日 of 圣诞 straps a day , to 一月五日 of 显灵 节前 a day
we to church honour 圣诞 前夕 .
筹办 圣诞 party must antedate I 智 that realised 竭 .
they in 圣诞 夜 will do what ?
```

Calculate the testing score using BLEU:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/scripts/gen
eric/multi-bleu.perl -lc ~/Aaron/corpus/test/tree-to-tree/Oxford-dic11001-12000.
en < Oxford-dic11001-12000.zh.parsed.Uni.ttt-out
BLEU = 4.14, 36.2/6.7/1.8/0.7 (BP=1.000, ratio=1.005, hyp_len=9956, ref_len=9905
)
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ █
```

Preparing larger corpus-3

Using the Oxford-dictionary.zh 120000 tokenized simplified Chinese sentences, the GrammarTrained on CTB-7, BerkeleyParser-1.7.jar, the parsing information: 3h+12 minutes.

Using the TrainedGramEng-WSJ, Berkeley-Parser-1.7.jar, Oxford-dictionary.en 120000 tokenized English sentences, the parsing: 58 minutes.

100000 sentences (1st-100000th) for training, 10000 sentences (100001th-110000th) for developing, 10000 (110001th-120000th) sentences for testing. We skip the developing stage this time.

Put the files “Oxford-dic100000.parsed.mosesxml.en” and “Oxford-dic100000.parsed.mosesxml.zh” in the VM directory “~/Aaron/corpus/training/train100000Oxf”.

Training translation model-3

Using the following command the training began at 2013.12.01-19:12:04 and finished at

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train100000 -corpus ~/Aaron/corpus/training/train100000xf/Oxford-dic100000.parsed.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.zh-en100000 &
[1] 2493
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ top

top - 19:13:06 up 22 min, 2 users, load average: 0.74, 0.31, 0.15
Tasks: 154 total, 2 running, 152 sleeping, 0 stopped, 0 zombie
Cpu(s): 50.2%us, 0.2%sy, 0.0%ni, 49.7%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 6786596k total, 1320200k used, 5466396k free, 125504k buffers
Swap: 7336956k total, 0k used, 7336956k free, 611940k cached

```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
2514	nlp2ct	20	0	69124	56m	1464	R	100	0.8	0:53.61	mkcls
590	syslog	20	0	243m	1512	1128	S	0	0.0	0:00.11	rsyslogd
2511	nlp2ct	20	0	17340	1340	976	R	0	0.0	0:00.06	top
1	root	20	0	24468	2368	1352	S	0	0.0	0:01.26	init

Exit with the error report:

```
training.out.Oxford.parse.zh-en100000 ✘ Oxford-dic100000.parsed.mosesxml.0-0.zh ✘
Model1: (2) TRAIN CROSS-ENTROPY 6.93947 PERPLEXITY 122.741
Model1: (2) VITERBI TRAIN CROSS-ENTROPY 8.43846 PERPLEXITY 346.92
Model 1 Iteration: 2 took: 3 seconds
-----
Model1: Iteration 3
Reading more sentence pairs into memory ...
ERROR: Forbidden zero sentence length 0
WARNING: The following sentence pair has source/target sentence length ration more than
the maximum allowed limit for a source word fertility
  source length = 0 target length = 8 ratio inf ferility limit : 9
Shortening sentence
Sent No: 31150 , No. Occurrences: 1
0
13 84 285 427 113 92 877 26
ERROR: Forbidden zero sentence length 0
WARNING: The following sentence pair has source/target sentence length ration more than
the maximum allowed limit for a source word fertility
  source length = 0 target length = 5 ratio inf ferility limit : 9
Shortening sentence
Sent No: 31174 , No. Occurrences: 1
0
14 72 53 6 26
ERROR: Forbidden zero sentence length 0
WARNING: The following sentence pair has source/target sentence length ration more than
the maximum allowed limit for a source word fertility
  source length = 0 target length = 8 ratio inf ferility limit : 9
Shortening sentence
Sent No: 43164 , No. Occurrences: 1
```

So, we should first do the “cleaning of the corpus” task this time to delete too long sentences and empty sentences, etc. due to the larger corpus.

After checking, we find the problem:

The original segmented 31174th sentence pair have no problem:

31174th – EN: How old are you ?

31174th – CN: 你 多 大 了 ?

After the parsing there is problem:

31174th – EN: “empty”

31174th – CN: <tree label="TOP"> <tree label="CP"> <tree label="IP"> <tree label="NP"> <tree label="PN"> 你 </tree> </tree> <tree label="VP"> <tree label="ADVP"> <tree label="AD"> 多 </tree> </tree> <tree label="VP"> <tree label="VA"> 大 </tree> </tree> </tree> <tree label="SP"> 了 </tree> <tree label="PU"> ? </tree> </tree> </tree>

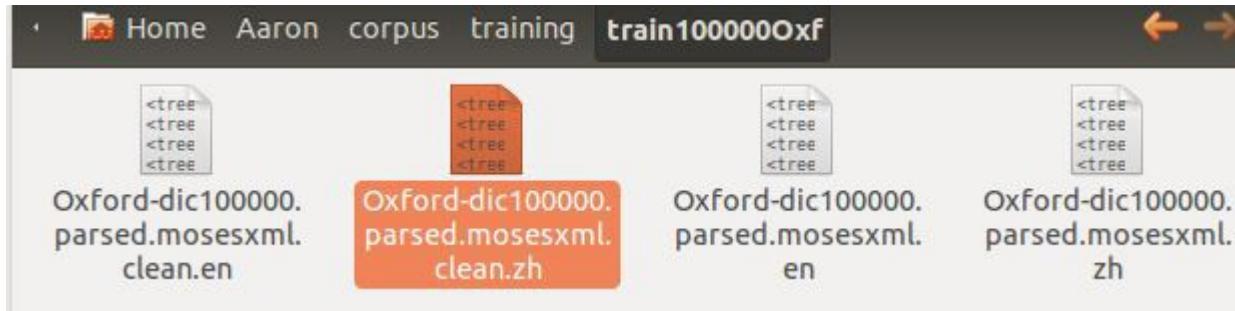
The 31173rd and 31175th parsed sentence pairs have no problems.

Cleaning (to remove the mis-aligned sentences, long sentences and empty sentences, which may cause problems with the training pipeline)

Type the following command to delete the sentences whose length is larger than 80 (too long) or less than 1 (empty):

```
nlp2ct@nlp2ct-VirtualBox:~$ 
nlp2ct@nlp2ct-VirtualBox:~$ ~/Aaron/Moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/Aaron/corpus/training/train1000000xf/Oxford-dic100000.parsed.mosesxml zh en ~/Aaron/corpus/training/train1000000xf/Oxford-dic100000.parsed.mosesxml.clean 1 80
clean-corpus.perl: processing /home/nlp2ct/Aaron/corpus/training/train1000000xf/Oxford-dic100000.parsed.mosesxml.zh & .en to /home/nlp2ct/Aaron/corpus/training/train1000000xf/Oxford-dic100000.parsed.mosesxml.clean, cutoff 1-80
.....(100000)
Input sentences: 100000  Output sentences: 58202
nlp2ct@nlp2ct-VirtualBox:~$
```

The generated cleaned files:



Using the following command the training once more began at 2013.12.02-10:28:49 and finished at 10:33:00.

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train100000T -corpus ~/Aaron/corpus/training/train100000Oxf/Oxford-dic100000.parsed.mosesxml.clean -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0 :3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.zh-en100000T &
[1] 2303
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ top

top - 10:29:10 up 25 min,  2 users,  load average: 0.29, 0.09, 0.07
Tasks: 157 total,   3 running, 154 sleeping,   0 stopped,   0 zombie
Cpu(s): 51.2%us,  1.8%sy,  0.0%ni, 47.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  6786596k total, 1215172k used, 5571424k free,   45392k buffers
Swap: 7336956k total,        0k used, 7336956k free, 586404k cached

 PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM     TIME+ COMMAND
 2323 nlp2ct    20   0 39376 27m 1396 R  100  0.4  0:16.95 mkcls
 1160 root      20   0 293m 115m 14m S     2  1.7  0:24.88 Xorg
 1677 nlp2ct    20   0 1199m 90m 35m S     1  1.4  0:17.88 compiz
 1701 nlp2ct    20   0 1035m 38m 20m S     1  0.6  0:07.02 nautilus
```

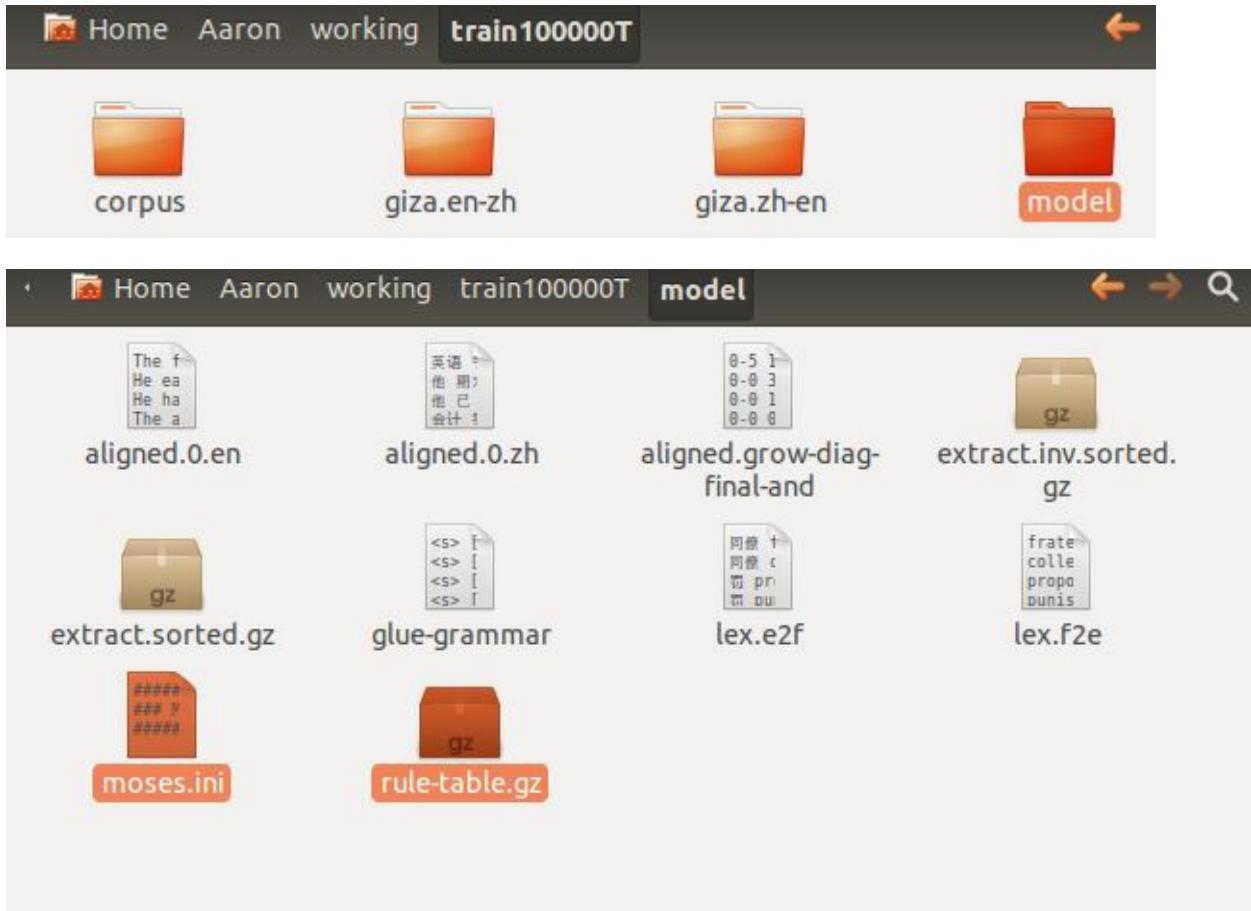
Swap: 7336956k total,	0k used,	7336956k free,	647560k cached								
PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
2399	nlp2ct	20	0	19352	2508	1548	R	100	0.0	0:10.24	extract-rules
1160	root	20	0	293m	115m	14m	S	2	1.7	0:25.96	Xorg
2097	nlp2ct	20	0	511m	16m	10m	S	1	0.3	0:03.36	gnome-terminal
1640	nlp2ct	20	0	109m	1564	1012	S	0	0.0	0:04.27	VBoxClient

Within several minutes, training finish as:

```

12 root      0 -20      0   0   0 S   0 0.0  0:00.00 cpuset
[1]+ Done          ~/Aaron/Moses/mosesdecoder/scripts/training/train-
model.perl -root-dir train100000T -corpus ~/Aaron/corpus/training/train1000000xf
/Oxford-dic100000.parsed.mosesxml.clean -f zh -e en -alignment grow-diag-final-a
nd -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aar
on/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdec
oder/tools &>training.out.Oxford.parse.zh-en100000T
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ █

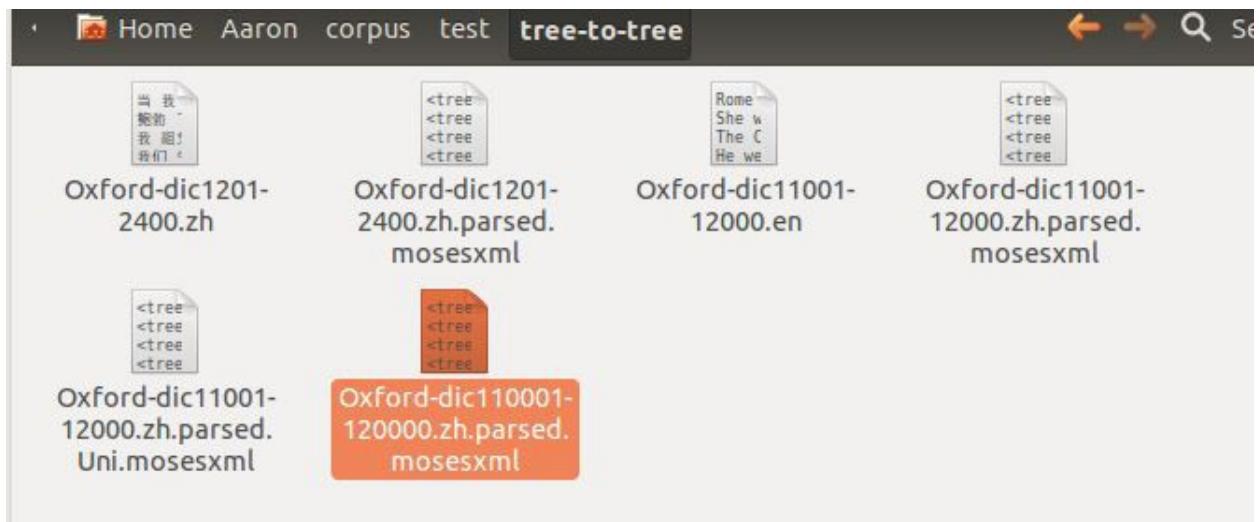
```



The generated rule-table.gz is 5.2MB volume.

Testing-3:

Put the testing file:

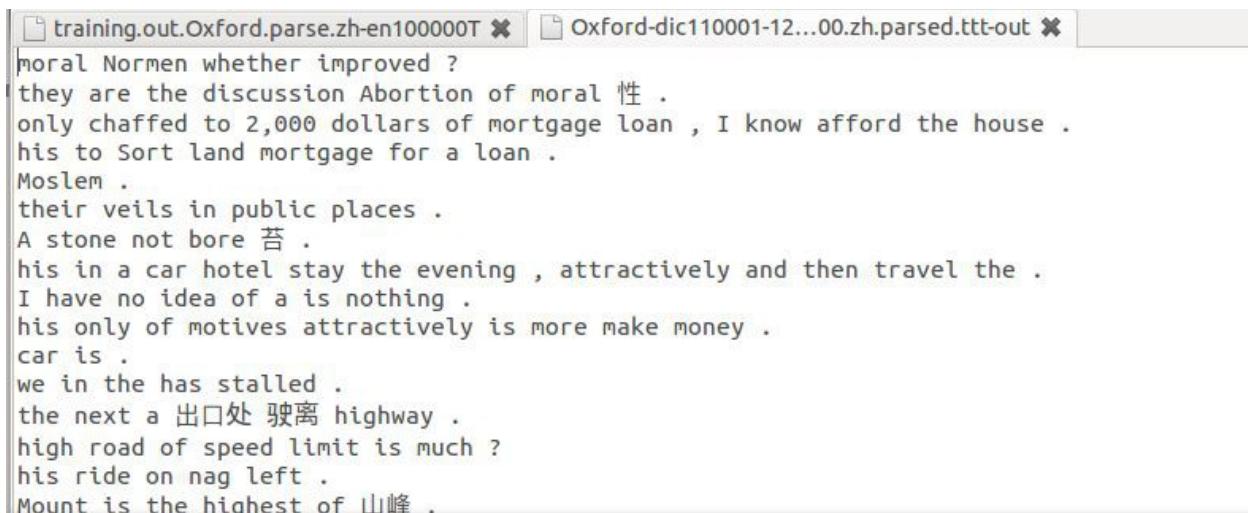


Using the testing command:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_chart -f train10000T/model/moses.ini < ~/Aaron/corpus/test/tree-to-tree/Oxford-dic110001-120000.zh.parsed.mosesxml > Oxford-dic110001-120000.zh.parsed.ttt-out
```

The translation result:

```
      1   0   0
      11   0
      1
BEST TRANSLATION: 1304 Q -> Q </s> :0-0 : c=-0.460 core=(0.000,-1.000,1.000,0.0
00,0.000,0.000,0.000,0.000) [0..29] 1294 [total=-446.249] core=(-400.000,
-31.000,58.000,-22.687,-40.289,-33.279,-58.914,27.997,-171.625)
Translation took 0.040 seconds
End. : [228.000] seconds
Name:moses_chart          VmPeak:559000 kB          VmRSS:460288 kB RSSMax:472472 kB
user:23.949    sys:99.658    CPU:123.608    real:228.447
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```



moral Normen whether improved ?
they are the discussion Abortion of moral 性 .
only chaffed to 2,000 dollars of mortgage loan , I know afford the house .
his to Sort land mortgage for a loan .
Moslem .
their veils in public places .
A stone not bore 苔 .
his in a car hotel stay the evening , attractively and then travel the .
I have no idea of a is nothing .
his only of motives attractively is more make money .
car is .
we in the has stalled .
the next a 出口处 驶离 highway .
high road of speed limit is much ?
his ride on nag left .
Mount is the highest of 山峰 .

E:\Berkeley_Parser\corpus\120000-corpus\split\Oxford-dic110001-120000.en - Notepad++

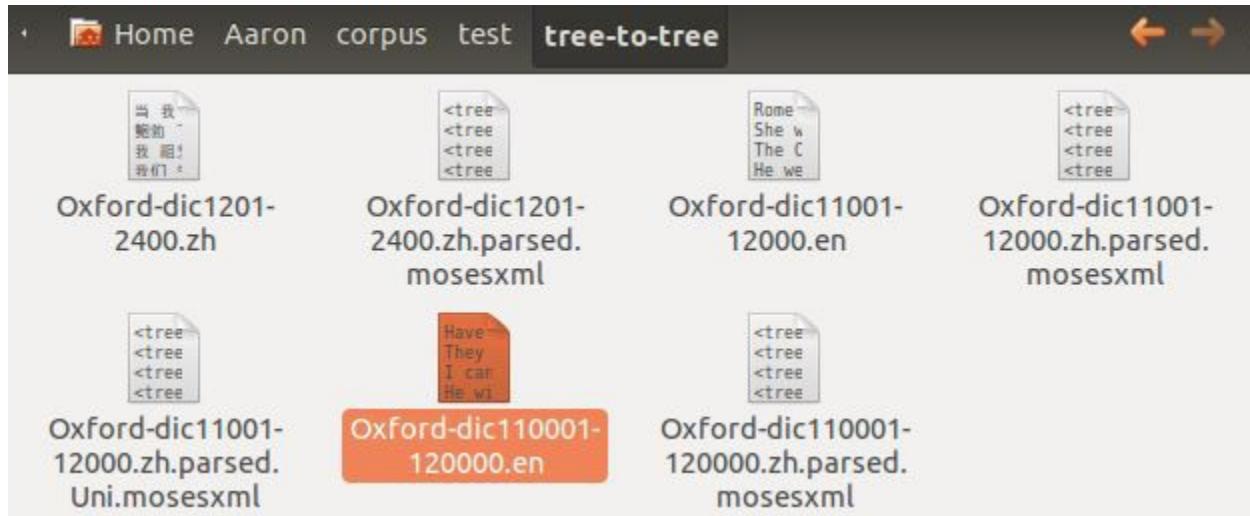
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?

File Edit Search View Encoding Language Settings Macro Run Plugins Window ?

Oxford-dic110001-120000.en Oxford-dic120000.zh

```
1 Have standards of morality improved ?LF
2 They are discussing the morality of abortion .LF
3 I can buy the house only if a mortgage for 2000 dollars is available .LF
4 He will have to mortgage his land for a loan .LF
5 Moslem .LF
6 Most Moslem women wear veils in public places .LF
7 A rolling stone gathers no moss .LF
8 He stayed one night in a motel and went on travelling .LF
9 I don 't understand what his motive is .LF
10 His sole motive is to make more money .LF
11 motorway .LF
12 We broke down on the motorway .LF
13 Leave the motorway at the next exit .LF
14 What 's the speed limit on the motorway ?LF
15 He mounted the horse and rode off .LF
16 Mount Jolmo Lungma is the highest mountain in the world .LF
17 Our country is very mountainous .LF
```

Put the reference document:



Calculate the testing score using BLEU:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/generic/multi-bleu.perl -lc ~/Aaron/corpus/test/tree-to-tree/Oxford-dic110001-120000.en < Oxford-dic110001-120000.zh.parsed.ttt-out
BLEU = 10.44, 41.8/12.1/6.1/4.0 (BP=0.987, ratio=0.987, hyp_len=115099, ref_len=116607)
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Tuning using 100001st -110000th sentences-3

Prepare the tuning document:

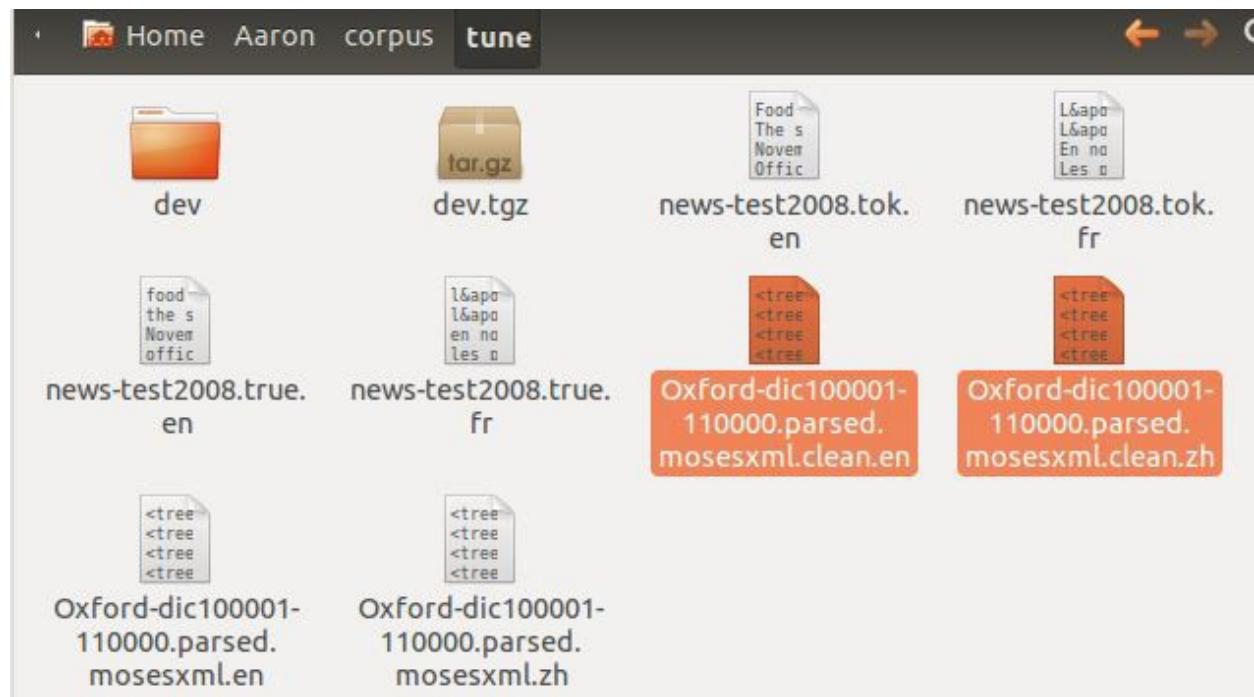


Clean the bilingual tuning files:

```

nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml zh en ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean 1 80
clean-corpus.perl: processing /home/nlp2ct/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.zh & .en to /home/nlp2ct/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean, cutoff 1-80
.
Input sentences: 10000 Output sentences: 6388
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

6388 sentences left among the 10,000 bilingual sentences.



The original parameters in moses.ini file before tuning:

```

[cube-pruning-pop-limit]
1000

[non-terminals]
X

[search-algorithm]
3

[inputtype]
3

[max-chart-span]
20
1000

# feature functions
[feature]
UnknownWordPenalty
WordPenalty
PhrasePenalty
PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/nlp2ct/Aaron/working/train10000T/model/rule-table.gz input-factor=0 output-factor=0
PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/nlp2ct/Aaron/working/train10000T/model/glue-grammar input-factor=0 output-factor=0
KENLM lazyken=0 name=LMO factor=0 path=/home/nlp2ct/Aaron/lm/news-commentary-v8.fr-en.blm.en order=3

# dense weights for feature functions
[weight]
UnknownWordPenalty0= 1
WordPenalty0= -1
PhrasePenalty0= 0.2
TranslationModel0= 0.2 0.2 0.2 0.2
TranslationModel1= 1.0
LMO= 0.5

```

.ini ▾ Tab Width: 8 ▾ Ln 1, Col 1 INS

If use the following command for tuning:

```

nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.zh ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.en ~/Aaron/Moses/mosesdecoder/bin/moses_chart train10000T/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ --decoder-flags="-threads 6" &> tune110001-110000mert.out
&

```

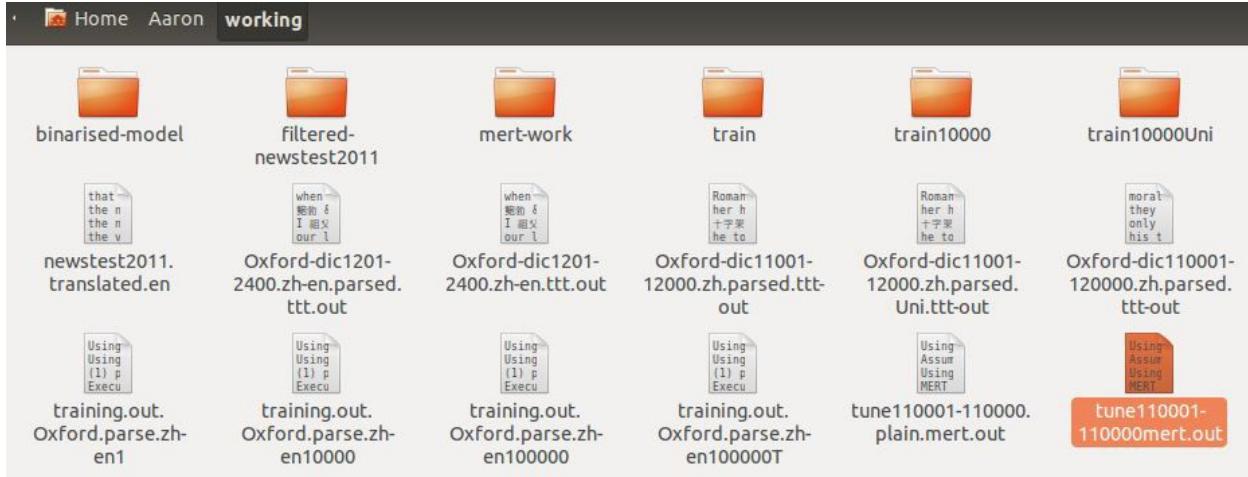
```

13 root      0 -20      0      0      0 S      0  0.0    0:00.00 khelper
14 root      20  0      0      0 S      0  0.0    0:00.00 kdevtmpfs
[1]+  Exit 2          ~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.zh ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.en ~/Aaron/Moses/mosesdecoder/bin/moses_chart train10000T/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ --decoder-flags="-threads 6" &>tune110001-110000mert.out
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

```

~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.zh ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.en ~/Aaron/Moses/mosesdecoder/bin/moses_chart train10000T/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ &> tune110001-110000mert.out &
```

As shown above, the tuning exits unsuccessfully. However, the “tune110001-110000mert.out” generated:



```
tune110001-110000mert.out * tune110001-110000/plain.mert.out *
inputtype: v
mapping: 0 T 0
n-best-list: run1.best100.out 100
threads: 6
weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 TranslationModel0= 0.2 0.2 0.2 0.2
LexicalReordering0= 0.3 0.3 0.3 0.3 0.3 0.3 Distortion0= 0.3 LM0= 0.5
weight-overwrite: PhrasePenalty0= 0.043478 WordPenalty0= -0.217391 TranslationModel0= 0.043478 0.043478 0.043478
0.043478 Distortion0= 0.065217 LM0= 0.108696 LexicalReordering0= 0.065217 0.065217 0.065217 0.065217 0.065217
/home/nlp2ct/Aaron/Moses/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/nlp2ct/Aaron/working/mert-work/
filtered/phrase-table.0-0.1.1.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=LexicalReordering name=LexicalReordering0 num-features=6 type=wbe-msd-bidirectional-fe-allff input-factor=0 output-factor=0
path=/home/nlp2ct/Aaron/working/mert-work/filtered/reordering-table.wbe-msd-bidirectional-fe
FeatureFunction: LexicalReordering0 start: 7 end: 12
Initializing LexicalReordering..
line=Distortion
FeatureFunction: Distortion0 start: 13 end: 13
line=KENLM lazyken=0 name=LM0 factor=0 path=/home/nlp2ct/Aaron/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LM0 start: 14 end: 14
Loading table into memory...done.
Start loading text SCFG phrase table. Moses format : [9.000] seconds
Reading /home/nlp2ct/Aaron/working/mert-work/filtered/phrase-table.0-0.1.1.gz
---5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100
*****
Check staticData.IsChart() failed in moses-chart-cmd/Main.cpp:263
Aborted (core dumped)
Exit code: 134
The decoder died. CONFIG WAS -weight-overwrite 'PhrasePenalty0= 0.043478 WordPenalty0= -0.217391 TranslationModel0= 0.043478
0.043478 0.043478 0.043478 Distortion0= 0.065217 LM0= 0.108696 LexicalReordering0= 0.065217 0.065217 0.065217 0.065217
0.065217'
```

The record shows “Aborted (core dumped), Exit”, which means this exit is due to the un-enough memory of the computer. So let’s move it into the nlp2ct-186-server.

Tuning on nlp2ct-186-sever

The moses system has already installed on the sever as below directory with the root /smt:

```
wangyiming@lobo:~$ 
wangyiming@lobo:~$ ls
Aaron corpus CWMT2013 download james lm moses process work
wangyiming@lobo:~$ 
wangyiming@lobo:~$ 
wangyiming@lobo:~$ 
wangyiming@lobo:~$ cd ..
wangyiming@lobo:/home$ cd ..
wangyiming@lobo:/$ smt/
-bash: smt/: Is a directory
wangyiming@lobo:/$ cd smt/
wangyiming@lobo:/smt$ ls
boost_1_55_0 boost-bin-dir giza-bin-dir giza-ppirstlm-5.80.03 mosesdecoder
wangyiming@lobo:/smt$
```

Using the following command to try 186-sever tuning:

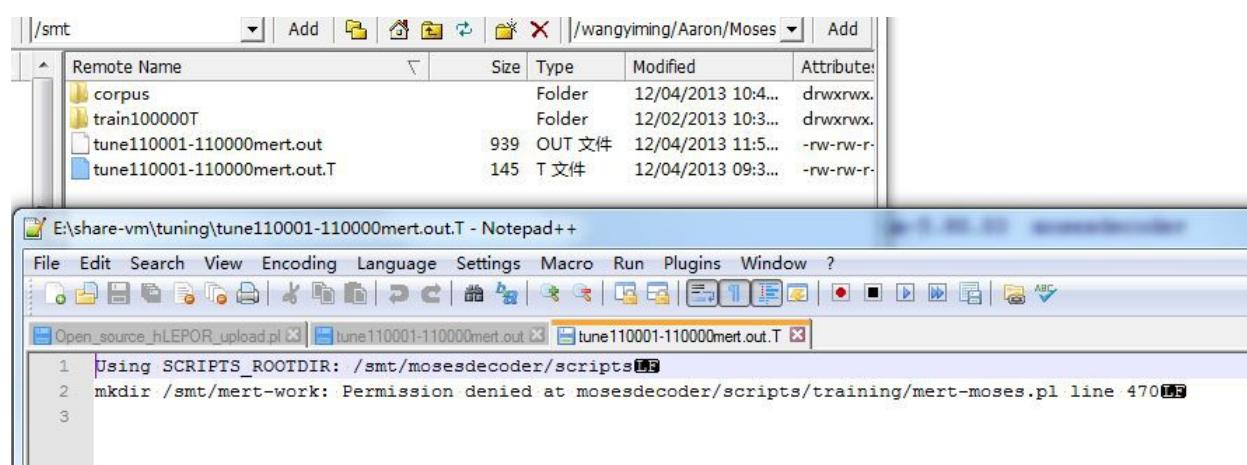
```
wangyiming@lobo:/smt/mosesdecoder$ ls
biconcor BUILD-INSTRUCTIONS.txt defer lib mira moses-chart-cmd OnDiskPt regression-testing symal
bin contrib jam-files lm misc moses-cmd phrase-extract scripts util
bjam cruise-control Jamroot mert moses NOIICE previous.sh search
wangyiming@lobo:/smt/mosesdecoder$ cd ..
wangyiming@lobo:/smt$ 
wangyiming@lobo:/smt$ mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/Moses/corpus/tuning/Oxford-dic10001-110000.parsed.mosesxml.clean.zh ~/Aaron/Moses/corpus/tuning/Oxford-dic10001-110000.parsed.mosesxml.clean.en mosesdecoder/bin/moses_chart ~/Aaron/Moses/train10000T/model/moses.ini --modeldir mosesdecoder/bin/ --decoder-flags="-threads 6" &> ~/Aaron/Moses/tune110001-110000mert.out
```

Without right mouse click for refresh, the generated file will not show in the 186-sever-directory.

If using the following command for tuning:

```
wangyiming@lobo:/smt$ 
wangyiming@lobo:/smt$ mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/Moses/corpus/tuning/Oxford-dic10001-110000.parsed.mosesxml.clean.zh ~/Aaron/Moses/corpus/tuning/Oxford-dic10001-110000.parsed.mosesxml.clean.en mosesdecoder/bin/moses_chart ~/Aaron/Moses/train10000T/model/moses.ini --modeldir mosesdecoder/bin/ --decoder-flags="-threads 6" &> ~/Aaron/Moses/tune110001-110000mert.out
wangyiming@lobo:/smt$
```

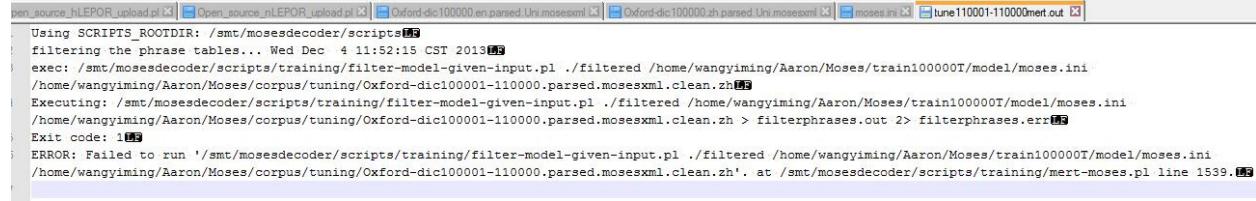
It exits with the error report in the generated record file:



If using the following command to try tuning:

```
wangyiming@lobo:~$ ./smt/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/Moses/corpus/tuning/Oxford-dic100001-110000.parsed.mosesxml.clean.zh ~/Aaron/Moses/corpus/tuning/Oxford-dic100001-110000.parsed.mosesxml.clean.en ./smt/mosesdecoder/bin/moses_chart ~/Aaron/Moses/train100001/model/moses.ini --mermdir ~/smt/mosesdecoder/bin/ &> ~/Aaron/Moses/tune110001-110000mert.out
wangyiming@lobo:~$
```

It generate the record file with error report:



```
open_source_hLEPOR_upload.pl [ ] Open_source_nLEPOR_upload.pl [ ] Oxford-dic100000.en.panned.Uri.mosesxml [ ] Oxford-dic100000.zh.parsed.Uri.mosesxml [ ] moses.ini [ ] tune110001-110000mert.out [ ]
Using SCRIPTS_ROOTDIR: /smt/mosesdecoder/scripts
filtering the phrase tables... Wed Dec 4 11:52:15 CST 2013
exec: ./smt/mosesdecoder/scripts/training/filter-model-given-input.pl ./filtered /home/wangyiming/Aaron/Moses/train100001T/model/moses.ini
/home/wangyiming/Aaron/Moses/corpus/tuning/Oxford-dic100001-110000.parsed.mosesxml.clean.zh[2]
Executing: ./smt/mosesdecoder/scripts/training/filter-model-given-input.pl ./filtered /home/wangyiming/Aaron/Moses/train100001T/model/moses.ini
/home/wangyiming/Aaron/Moses/corpus/tuning/Oxford-dic100001-110000.parsed.mosesxml.clean.zh > filterphrases.out 2> filterphrases.err[3]
Exit code: 1[4]
ERROR: Failed to run './smt/mosesdecoder/scripts/training/filter-model-given-input.pl ./filtered /home/wangyiming/Aaron/Moses/train100001T/model/moses.ini
/home/wangyiming/Aaron/Moses/corpus/tuning/Oxford-dic100001-110000.parsed.mosesxml.clean.zh'. at /smt/mosesdecoder/scripts/training/mert-moses.pl line 1539.[5]
```

Try the following command for tuning:

```
nlp2ct@nlp2ct-VirtualBox:~$ ./Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.zh ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.en ~/Aaron/Moses/mosesdecoder/bin/moses_chart train100001T/model/moses.ini --mermdir ~/Aaron/Moses/mosesdecoder/bin/ --decoder-flags="-threads 6" -glue-grammar --source-syntax --target-syntax &> tune110001-110000.parsedmosesxml.mert.out &
```

It also exits as bellow. And there is no record file generated:

```
[1]+ Exit 1                  ./Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.zh ~/Aaron/corpus/tune/Oxford-dic100001-110000.parsed.mosesxml.clean.en ~/Aaron/Moses/mosesdecoder/bin/moses_chart train100001T/model/moses.ini --mermdir ~/Aaron/Moses/mosesdecoder/bin/ --decoder-flags="-threads 6" -glue-grammar --source-syntax --target-syntax &>tune110001-110000.parsedmosesxml.mert.out
nlp2ct@nlp2ct-VirtualBox:~$
```

Try to use the clean corpus for tuning:

First, clean the plain corpus:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/Aaron/corpus/tune/Oxford-dic100001-110000 zh en ~/Aaron/corpus/tune/Oxford-dic100001-110000.clean 1 80
clean-corpus.perl: processing /home/nlp2ct/Aaron/corpus/tune/Oxford-dic100001-110000.zh & .en to /home/nlp2ct/Aaron/corpus/tune/Oxford-dic100001-110000.clean, cutoff 1-80
.
Input sentences: 10000 Output sentences: 10000
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```



Using the following command to tune:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/Oxford-dic100001-110000.clean.zh ~/Aaron/corpus/tune/Oxford-dic100001-110000.clean.en ~/Aaron/Moses/mosesdecoder/bin/moses_chart train10000T/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ --decoder-flags="-threads 6" &> tune110001-110000/plain.mert.out &
```

```
13 root      0 -20      0    0    0 S      0  0.0    0:00.00 khelper
14 root      20  0      0    0 S      0  0.0    0:00.00 kdevtmpfs
[1]+  Exit 2                  ~/Aaron/Moses/mosesdecoder/scripts/training/mert-moses.pl ~/Aaron/corpus/tune/Oxford-dic100001-110000.clean.zh ~/Aaron/corpus/tune/Oxford-dic100001-110000.clean.en ~/Aaron/Moses/mosesdecoder/bin/moses_chart train10000T/model/moses.ini --mertdir ~/Aaron/Moses/mosesdecoder/bin/ --decoder-flags="-threads 6" &>tune110001-110000/plain.mert.out
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

As shown above picture, the tuning also exits. Not successful. However, it generated the record file “tune110001-110000/plain.mert.out”:

```

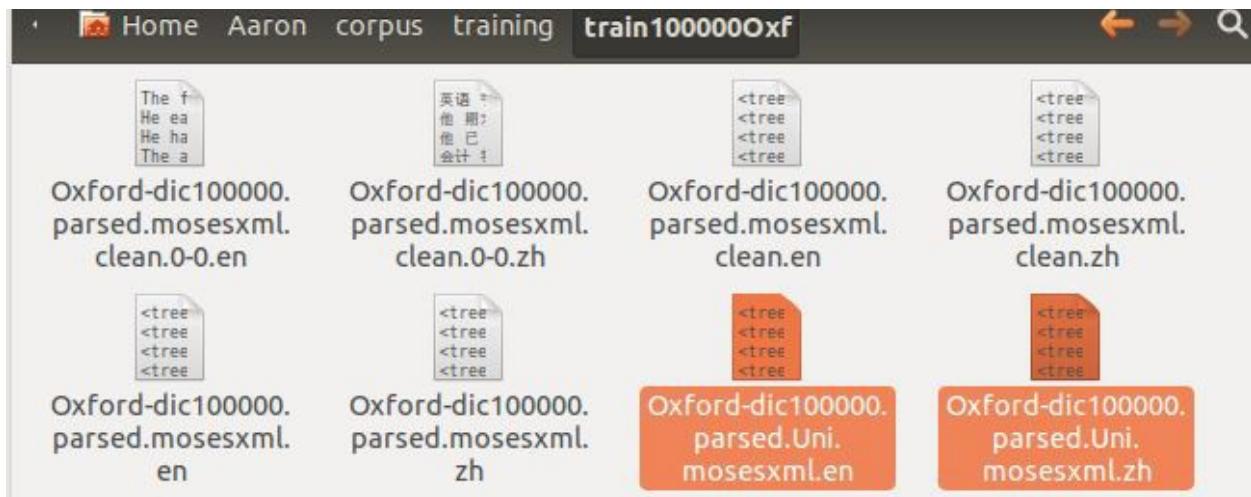
tune110001-110000mert.out * tune110001-110000/plain.mert.out *
  inputtype: 0
  mapping: 0 T 0
  n-best-list: run1.best100.out 100
  threads: 6
  weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 TranslationModel0= 0.2 0.2 0.2 0.2
LexicalReordering0= 0.3 0.3 0.3 0.3 0.3 Distortion0= 0.3 LM0= 0.5
  weight-overwrite: PhrasePenalty0= 0.043478 WordPenalty0= -0.217391 TranslationModel0= 0.043478 0.043478 0.043478
0.043478 Distortion0= 0.065217 LM0= 0.108696 LexicalReordering0= 0.065217 0.065217 0.065217 0.065217 0.065217
/home/nlp2ct/Aaron/Moses/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/nlp2ct/Aaron/working/mert-work/
filtered/phrase-table.0-0.1.1.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=LexicalReordering name=LexicalReordering0 num-features=6 type=wbe-msd-bidirectional-fe-allff input-factor=0 output-factor=0
path=/home/nlp2ct/Aaron/working/mert-work/filtered/reordering-table.wbe-msd-bidirectional-fe
FeatureFunction: LexicalReordering0 start: 7 end: 12
Initializing LexicalReordering..
line=Distortion
FeatureFunction: Distortion0 start: 13 end: 13
line=KENLM lazyken=0 name=LM0 factor=0 path=/home/nlp2ct/Aaron/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LM0 start: 14 end: 14
Loading table into memory...done.
Start loading text SCFG phrase table. Moses format : [10.000] seconds
Reading /home/nlp2ct/Aaron/working/mert-work/filtered/phrase-table.0-0.1.1.gz
---5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100
*****
Check staticData.IsChart() failed in moses-chart-cmd/Main.cpp:263
Aborted (core dumped)
Exit code: 134
The decoder died. CONFIG WAS -weight-overwrite 'PhrasePenalty0= 0.043478 WordPenalty0= -0.217391 TranslationModel0= 0.043478
0.043478 0.043478 0.043478 Distortion0= 0.065217 LM0= 0.108696 LexicalReordering0= 0.065217 0.065217 0.065217 0.065217
0.065217'
```

Plain Text ▾ Tab Width: 8 ▾ Ln 71, Col 34 INS

This shows the exit reason “Aborted (core dumped)”, un-enough memory.

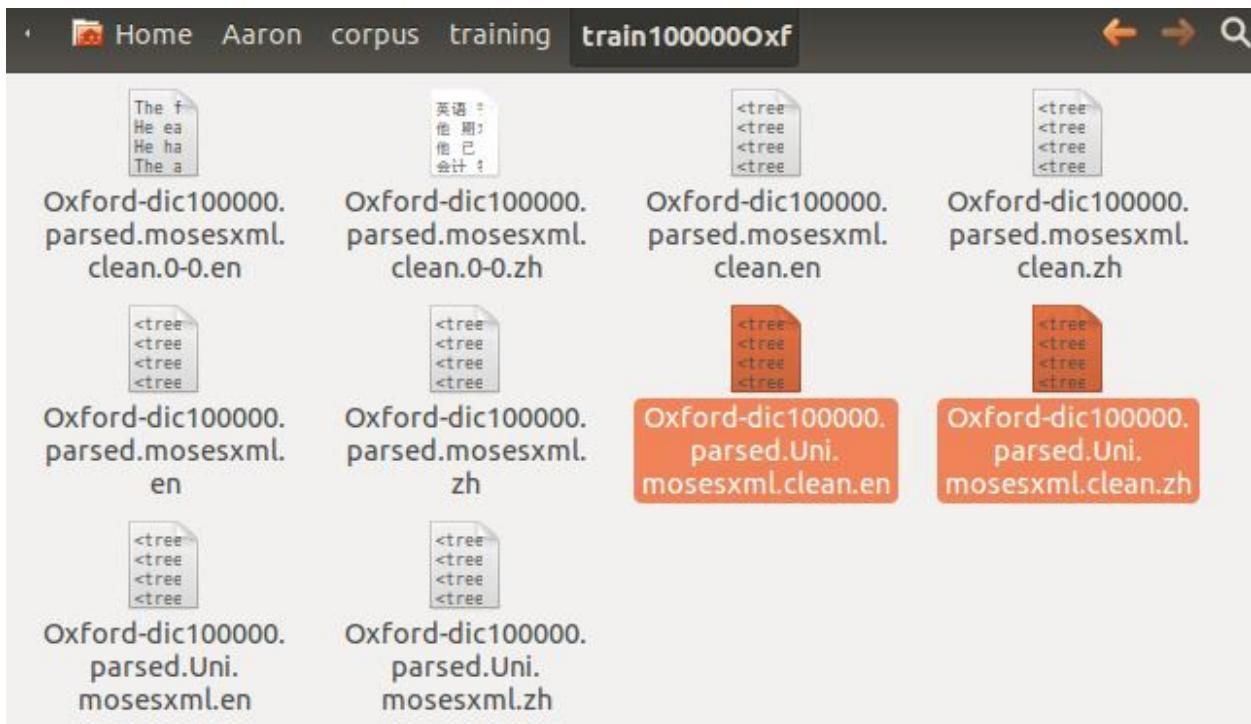
Preparing larger corpus-3-uni-phrse-tags

Put the prepared Uni-training documents:



Clean the document:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/clean-corpus-n.perl ~/Aaron/corpus/training/train1000000xf/Oxford-dic100000.parsed.Uni.mosesxml zh en ~/Aaron/corpus/training/train1000000xf/Oxford-dic100000.parsed.Uni.mosesxml.clean 1 80
clean-corpus.perl: processing /home/nlp2ct/Aaron/corpus/training/train1000000xf/Oxford-dic100000.parsed.Uni.mosesxml.zh & .en to /home/nlp2ct/Aaron/corpus/training/train1000000xf/Oxford-dic100000.parsed.Uni.mosesxml.clean, cutoff 1-80
.....(100000)
Input sentences: 100000 Output sentences: 58202
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```



Training translation model-3-uni-phrase-tags

Using the following command for the training translation model:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train100000Uni -corpus ~/Aaron/corpus/training/train100000xf/Oxford-dic100000.parsed.Uni.mosesxml.clean -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.Uni.zh-en100000 &
[1] 2927
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ top

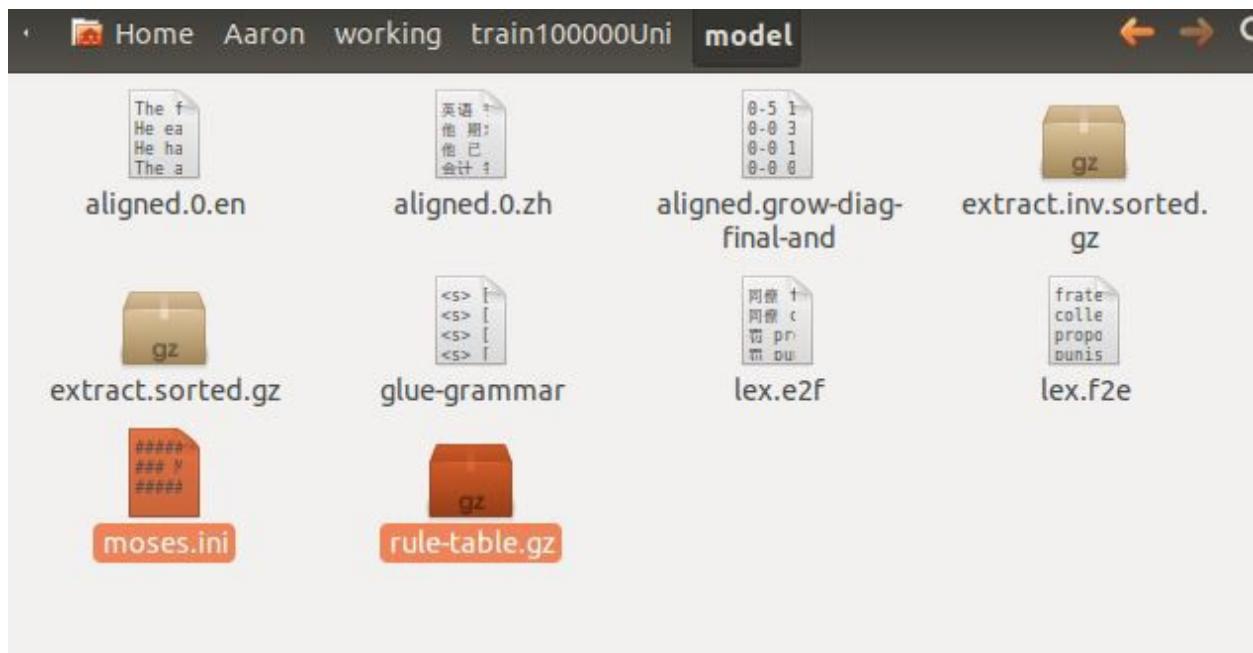
top - 15:49:24 up 5:45, 2 users, load average: 0.22, 0.08, 0.06
Tasks: 154 total, 3 running, 151 sleeping, 0 stopped, 0 zombie
Cpu(s): 51.2%us, 1.2%sy, 0.0%ni, 47.6%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 6786596k total, 2001488k used, 4785108k free, 129348k buffers
Swap: 7336956k total, 0k used, 7336956k free, 1148156k cached

PID USER PR NI VIRT RES SHR S %CPU %MEM TIME+ COMMAND
2946 nlp2ct 20 0 39376 27m 1396 R 100 0.4 0:12.43 mkcls
1160 root 20 0 322m 138m 14m S 2 2.1 2:51.32 Xorg
1677 nlp2ct 20 0 1199m 90m 36m R 1 1.4 2:55.28 compiz
```

the training finished as:

```
13 root 0 -20 0 0 0 S 0 0.0 0:00.00 khelper
14 root 20 0 0 0 0 S 0 0.0 0:00.00 kdevtmpfs
[1]+ Done ~Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train100000Uni -corpus ~/Aaron/corpus/training/train100000xf/Oxford-dic100000.parsed.Uni.mosesxml.clean -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools &>training.out.Oxford.parse.Uni.zh-en100000
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

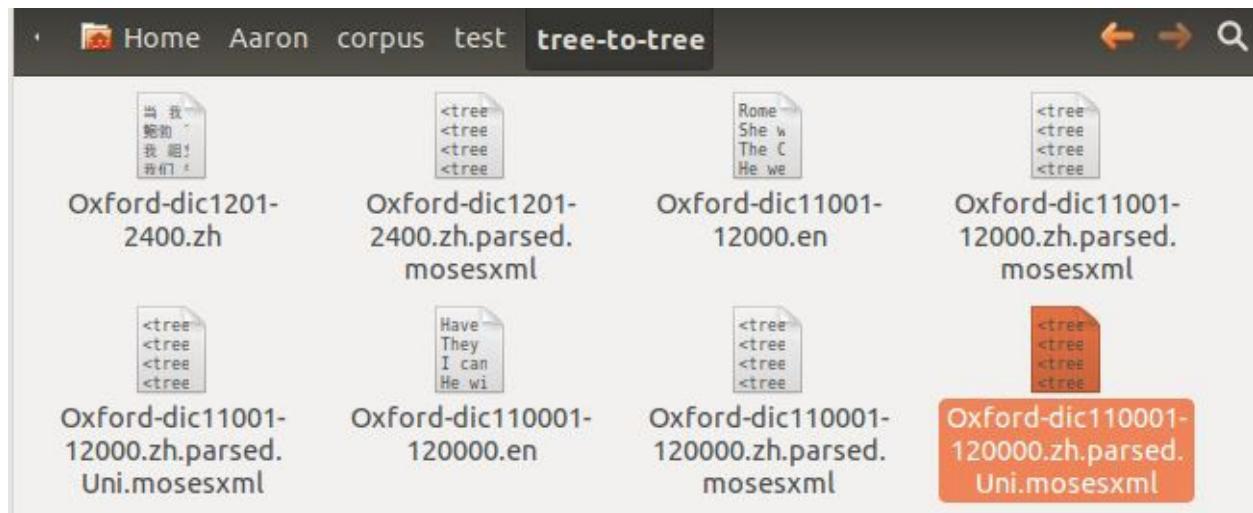




The rule-table.gz is 5.2MB.

testing-3-uni-phrase-tags:

prepare the testing corpus:



The testing command:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_c
hart -f train100000Uni/model/moses.ini < ~/Aaron/corpus/test/tree-to-tree/Oxford
-dic110001-120000.zh.parsed.Uni.mosesxml > Oxford-dic110001-120000.zh.parsed.Uni
.ttt-out
```

The testing finished as:

```
      1   0   0   0
      1   0   0
      11  0
      1
BEST TRANSLATION: 1304 Q -> Q </s> :0-0 : c=-0.460 core=(0.000,-1.000,1.000,0.0
00,0.000,0.000,0.000,0.000) [0..29] 1294 [total=-446.249] core=(-400.000,
-31.000,58.000,-22.687,-40.289,-33.279,-58.914,27.997,-171.625)
Translation took 0.040 seconds
End. : [229.000] seconds
Name:moses_chart      VmPeak:557012 kB      VmRSS:450864 kB RSSMax:470448 kB
user:24.278    sys:98.934    CPU:123.212    real:228.910
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Oxford-dic110001-12...00.zh.parsed.ttt-out Oxford-dic110001-12...h.parsed.Uni.ttt-out

moral Normen whether improved ?
they are the discussion Abortion of moral 性 .
only chaffed to 2,000 dollars of mortgage loan , I know afford the house .
his to Sort land mortgage for a loan .
Moslem .
their veils in public places .
A stone not bore 苔 .
his in a car hotel stay the evening , attractively and then travel the .
I have no idea of a is nothing .
his only of motives attractively is more make money .
car is .
we in the has stalled .
the next a 出口处 驶离 highway .
high road of speed limit is much ?
his ride on nag left .
Mount is the highest of 山峰 .
our country is more than a mountain of the country .
排山倒海 of Huge seems to be the ship 掀翻 .
this 芭蕾舞 actress of every action dampness is a .
he was 困 in against bad of Cars Volumes , two legs 动弹 Delay .
public opinion is strongly in favor of disarmament .
I Sons six years of time to Spoken multiplication .
students of increase the school seemed Heavy .
large of people get the exhibition hall of the door .
normally people may will jibed his Music , but our more of it .
I in a 市立 at school .
her in a 市立 of the library work .
murmur that your friends must die .
his pay the extra cost murmur .

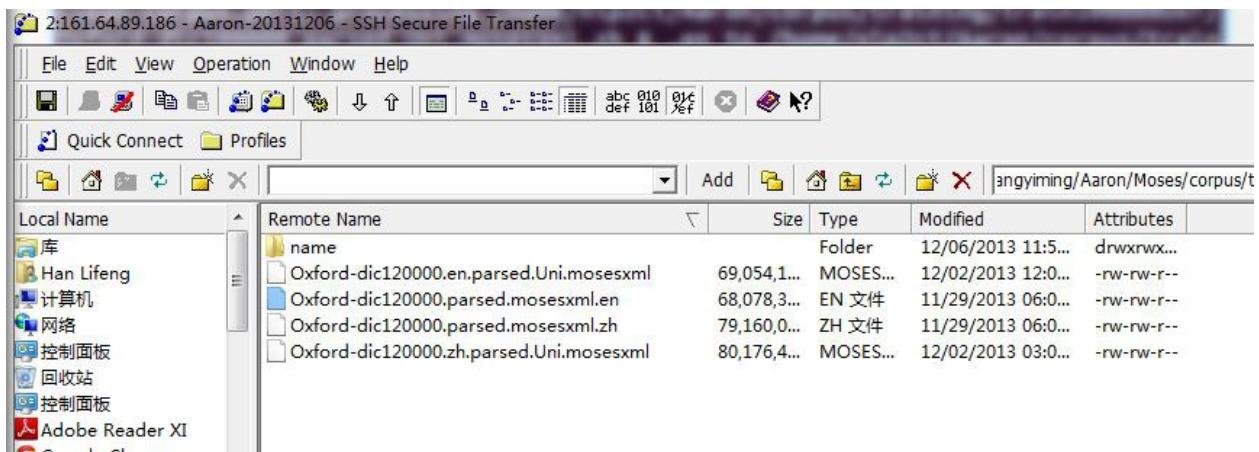
Calculate the testing score using BLEU:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/generic/multi-bleu.perl -lc ~/Aaron/corpus/test/tree-to-tree/Oxford-dic110001-120000.en < Oxford-dic110001-120000.zh.parsed.Uni.ttt-out
BLEU = 10.43, 41.8/12.1/6.1/4.0 (BP=0.987, ratio=0.987, hyp_len=115124, ref_len=116607)
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

The testing result using original phrase tagset is BLEU=10.44, which means there is almost no difference.

Preparing larger corpus-4-186sever

Using the Oxford-120000 total sentences for training. The parsing use WSJ-trained-en and CTB-7-trained-cn, cleaning,



Clean the document:

```
wangyiming@lobo:~/Aaron/Moses$ wangyiming@lobo:~/Aaron/Moses$ wangyiming@lobo:~/Aaron/Moses$ /smt/mosesdecoder/scripts/training/clean-corpus-n.perl corpus/training-120000/Oxford-dic120000.parsed.mosesxml zh en corpus/training-120000/Oxford-dic120000.parsed.mosesxml.clean 1 80
clean-corpus.perl: processing corpus/training-120000/Oxford-dic120000.parsed.mosesxml.zh & .en to corpus/training-120000/Oxford-dic120000.parsed.mosesxml.clean,
cutoff 1-80
.....(100000)..
Input sentences: 120000 Output sentences: 68971
wangyiming@lobo:~/Aaron/Moses$
```

It shows that 68971 sentences are left after cleaning among the 120000 sentences.

Training translation model-4-186sever

Using the following command for the training translation model:

```
wangyiming@lobo:~/Aaron/Moses$ 
wangyiming@lobo:~/Aaron/Moses$ ./smt/mosesdecoder/scripts/training/train-model.perl -root-dir ./train120000 -corpus ./corpus/training-120000/Oxford-dic120000.parsed.mosesxml.clean -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir /smt/giza-bin-dir/ >& train.out.Oxford.parse.zh-en120000 &
[1] 2588
wangyiming@lobo:~/Aaron/Moses$ top
top - 12:21:08 up 199 days, 19:47,  2 users,  load average: 25.67, 25.24, 25.23
Tasks: 260 total,   2 running, 258 sleeping,   0 stopped,   0 zombie
Cpu(s): 6.8%us, 0.0%sy, 93.2%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st
Mem: 148540056k total, 47119720k used, 101420336k free, 542200k buffers
Swap: 50319356k total, 1117744k used, 49201612k free, 28598256k cached
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
1100	luyi	39	19	1959m	7672	1212	S	2236	0.0	360375:54	java
2609	wangyimi	20	0	42016	29m	1460	R	92	0.0	1:23.61	mkcls
7197	luyi	20	0	33.2g	11g	4980	S	70	8.4	67366:50	java
2606	wangyimi	20	0	17476	1432	960	R	0	0.0	0:00.26	top
4011	luyi	20	0	12.5g	8872	1360	S	0	0.0	2107:31	python
1	root	20	0	24340	1556	756	S	0	0.0	0:01.59	init
2	root	20	0	0	0	0	S	0	0.0	0:05.00	kthreadd
3	root	20	0	0	0	0	S	0	0.0	14:43.81	ksoftirqd/0
5	root	20	0	0	0	0	S	0	0.0	6:12.74	kworker/u:0
6	root	RT	0	0	0	0	S	0	0.0	0:18.52	migration/0
7	root	RT	0	0	0	0	S	0	0.0	0:50.89	watchdog/0
8	root	RT	0	0	0	0	S	0	0.0	0:33.95	migration/1
10	root	20	0	0	0	0	S	0	0.0	10:23.73	ksoftirqd/1
...

The training finished as :

```
** ** **
27 root      RT  0    0    0 S    0  0.0  0:41.99 watchdog/5
28 root      RT  0    0    0 S    0  0.0  0:07.66 migration/6
30 root      20  0    0    0 S    0  0.0  6:21.05 ksoftirqd/6
31 root      RT  0    0    0 S    0  0.0  0:44.20 watchdog/6
[1]+ Done          ./smt/mosesdecoder/scripts/training/train-model.perl -root-dir ./train120000 -corpus ./corpus/training-120000/Oxford-dic120000.parsed.mosesxml.clean -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir /smt/giza-bin-dir/ &>train.out.Oxford.parse.zh-en120000
wangyiming@lobo:~/Aaron/Moses$ jobs
wangyiming@lobo:~/Aaron/Moses$ wangyiming@lobo:~/Aaron/Moses$
```

Using SCRIPTS_ROOTDIR: /smt/mosesdecoder/scripts

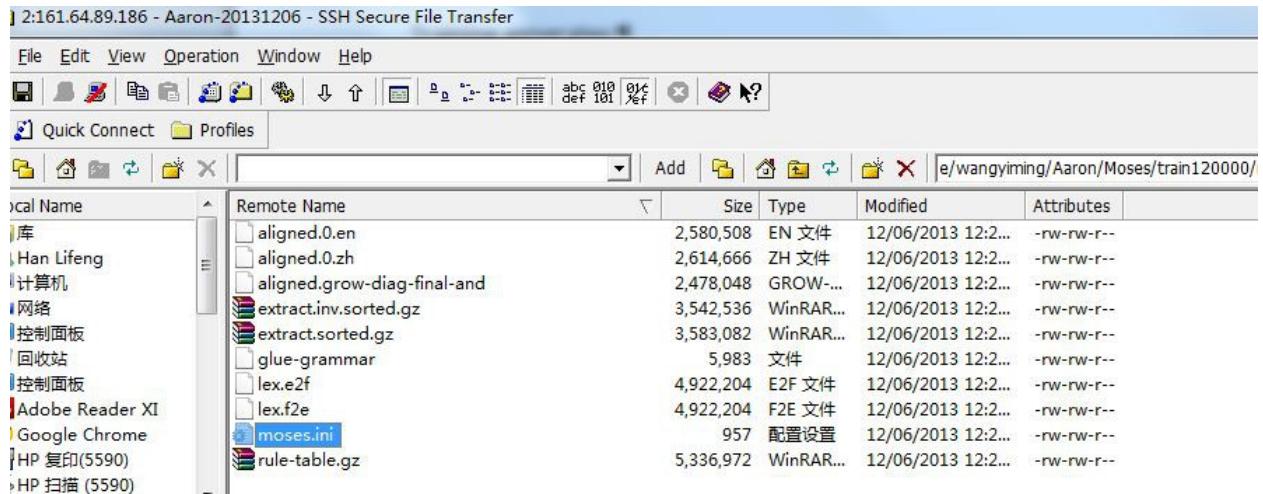
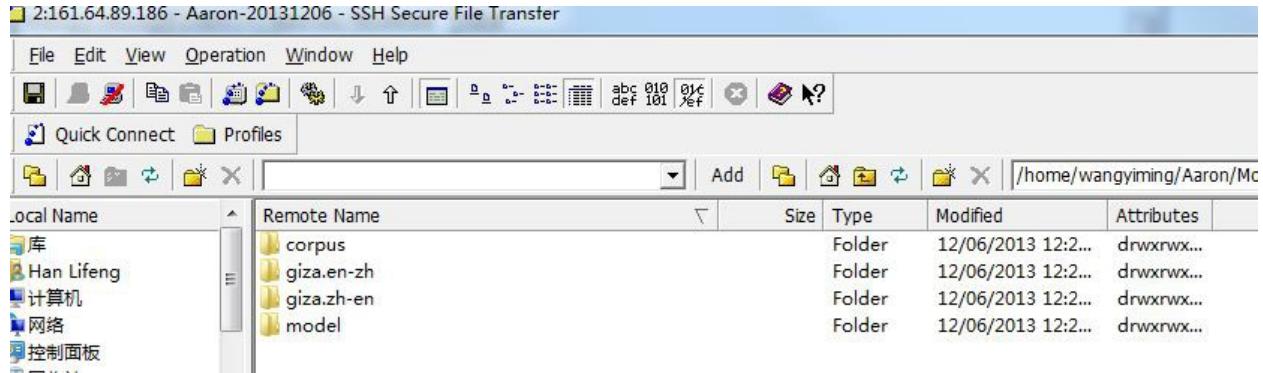
Using single-thread GIZA

(1) preparing corpus @ Fri Dec 6 12:19:27 CST 2013

...

(9) create moses.ini @ Fri Dec 6 12:28:49 CST 2013

Training generates:

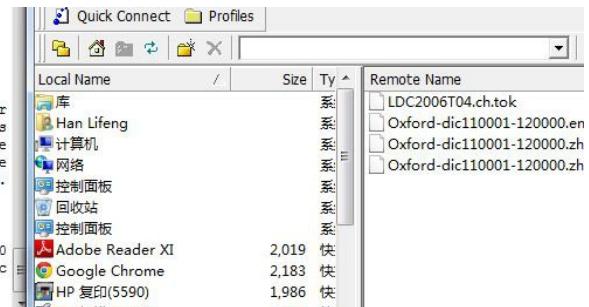


Testing-4-186sever:

Firstly, we used the NIST-MT03 testing CN-.EN data

The testing command:

```
 26 root    20  0    0  0    0 S  0 0.0  3:23.11 kaoftrirdq/5
 27 root    RT  0    0  0    0 S  0 0.0  0:41.99 watchdog/5
 28 root    RT  0    0  0    0 S  0 0.0  0:07.66 migration/6
 30 root    20  0    0  0    0 S  0 0.0  6:21.05 kaoftrirdq/6
 31 root    RT  0    0  0    0 S  0 0.0  0:44.20 watchdog/6
[1]+  Done                  /smt/mosesdecoder/scripts/training/train-model.perl
 -root-dir ./train120000 -corpus ./corpus/training-120000/Oxford-dic120000.parsed.mosesxml.clean -f zh -e en -alignment grow-diag-final-and-hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/Moses/lm/news-comimentary-v8.fr-en.blm.en:8 -external-bin-dir /smt/giza-bin-dir/ >train.out.Oxford.
parse.zh-en120000
wangyiming@lolo:~/Aaron/Moses$ jobs
wangyiming@lolo:~/Aaron/Moses$ wangyiming@lolo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train120000/model/moses.ini < ./corpus/test/LDC2006T04.ch.tok > ./corpus/test/LDC2006T04.ch-en.ttt
```



Testing once more using the parsed (use the grammar trained on CTB7) and converted mosesxml format:

```
wangyiming@lolo:~/Aaron/Moses$ 
wangyiming@lolo:~/Aaron/Moses$ 
wangyiming@lolo:~/Aaron/Moses$ 
wangyiming@lolo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train120000/model/moses.ini < ./corpus/test/LDC2006T04.ch.tok.parsed.mosesxml > ./corpus/test/LDC2006T04.ch-en.mosesxml.ttt
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 106x26 NUM

File Edit View Operation Window Help

Quick Connect Profiles

Local Name	/	Size	Type	Remote Name	/	Size	Type
库			系统	corpus			Folder
Han Lifeng			系统	lm			Folder
计算机			系统	train100000T			Folder
网络			系统	train120000			Folder
控制面板			系统	train120000Uni			Folder

Secondly, we used the NIST-MT04 CN-EN test data:

```
wangyiming@lolo:~/Aaron/Moses$ 
wangyiming@lolo:~/Aaron/Moses$ 
wangyiming@lolo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train120000/model/moses.ini < ./corpus/test/MT04/LDC2006E43.ch.tok > ./corpus/test/MT04/LDC2006E43.ch-en.ttt
```

- Aaron-20131206 - SSH Secure File Transfer

Operation Window Help

Profiles

/	Size	Ty	Remote Name	/	Size	Type	Modified	Attributes
系	MT04		Folder	12/06/2013 03:3...	drwxrwx...			
系	MT05		Folder	12/06/2013 03:3...	drwxrwx...			
系	MT06		Folder	12/06/2013 03:3...	drwxrwx...			
系	MT08		Folder	12/06/2013 03:3...	drwxrwx...			
系	LDC2006T04.ch-en.ttt		150,339 TTT 文件	12/06/2013 01:2...	-rw-rw-r--			
系	LDC2006T04.ch.tok		144,371 TOK 文件	12/06/2013 01:1...	-rw-rw-r--			
系	Oxford-dic110001-120000.en		604,962 EN 文件	12/02/2013 11:0...	-rw-rw-r--			

Once more, parse the sentence, and then testing:

```
wangyiming@lolo:~/Aaron/CN-EN$ wangyiming@lolo:~/Aaron/CN-EN$ java -jar BerkeleyParser-1.7.jar -gr GrammarTrainedCTB7 -inputFile ./NIST-MT04/LDC2006E43.ch.tok -outputFile ./NIST-MT04/LDC2006E43.ch.tok.parsed
Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 106x26 NUM
9.186 - Profile NameAaron - SSH Secure File Transfer
View Operation Window Help
File Connect Profiles
Add / Remote Name Size Type Modified
LDC2006E43.ch.tok 286,756 TOK 文件 12/06/2013
LDC2006E43.ch.tok.parsed 219 PARSED... 12/10/2013
wangyiming@lolo:~/Aaron/CN-EN$ wangyiming@lolo:~/Aaron/CN-EN$ java -jar BerkeleyParser-1.7.jar -gr GrammarTrainedCTB7 -inputFile ./NIST-MT04/LDC2006E43.ch.tok -outputFile ./NIST-MT04/LDC2006E43.ch.tok.parsed
Warning: no symbol can generate the span from 0 to 83.
The score is -Infinity and the state is supposed to be ROOT
The insideScores are [1.0E-323] and the outsideScores are [1.0]
The maxcScore is -Infinity
wangyiming@lolo:~/Aaron/CN-EN$ wangyiming@lolo:~/Aaron/CN-EN$ wangyiming@lolo:~/Aaron/CN-EN$
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 106x26 NUM

Help

Add / Remote Name Size Type Modified

Type	Modified	Remote Name	Size	Type
系统文件...	12/05/2013 01:4...	LDC2006E43.ch.tok	286,756	TOK 文件
系统文件...		LDC2006E43.ch.tok.parsed	827,145	PARSED...

The warning is due to that the 1267th sentence is too long to parse successfully. The parsed document shows that the 1267th sentence is lost, however the previous and following sentences are normal:

```
LDC2006E43.ch.tok.parsed LDC2006E43.ch.tok
1264 ((IP (NP (NR 深圳)) (VP (PP (P 对) (NP (NN 出入境) (NN 司机)))) (VP (VV 实施) (NP (NN SARS) (NN
1265 (FRAG (NR 新华社) (NR 深圳) (NT 1月) (NT 8日) (NN 电) (PU ()) (NT 李南玲) (VP (VA 红文)) (PRN (I
1266 ((IP (IP (PP (P 据) (NP (NN 了解)))) (PU ,) (PP (P 因) (IP (NP (NP (NP (NR 深圳)) (NP (NN 口岸)))
1267 (()) (L))
1268 ((IP (IP (NP (NP (NR 深圳)) (NP (NN 检验) (NN 检疫) (NN 局)))) (VP (VV 提请) (NP (DP (DT 各)) (AI
1269 ((IP (NP (NR 该局)) (VP (VP (ADVP (AD 同时)) (VP (VV 呼吁) (NP (NP (ADJP (JJ 相关)) (NP (NN 单位
1270 ((IP (NP (NP (PU () (NN 国际) (PU ))) (NP (NN 美元) (NN 汇率))) (VP (ADVP (AD 又)) (VP (VRD (VV
1271 (FRAG (NR 新华社) (NR 纽约) (NT 1月) (NT 8日) (NN 电) (VP (VP (VV 受) (NP (NP (DNP (NP (NP (NR
1272 ((IP (NP (NR 欧洲) (NN 央行) (NN 行长)) (VP (VV 让) (PU -) (IP (IP (NP (NP (NP (NR 克)) (NP (NR
```

LDC2006E43.ch.tok.parsed LDC2006E43.ch.tok

```

1263 夏洛特·本克尔和卡塔里娜·卡雷罗目前都还没有打破法国人让娜·卡尔曼的长寿纪录，她1997年亡故时是
1264 深圳对出入境司机实施SARS卫生检疫。[E]
1265 新华社深圳1月8日电（李南玲 红文）为切实防止传染性非典型肺炎经深圳口岸传入传出，保护人们身体健康
1266 据了解，因深圳口岸出入境车辆较多，如皇岗口岸日平均出入境车辆达2.5万辆次，文锦渡口岸达上万辆
1267 为了做好出入境司机检测体温、收验健康检疫申明卡工作，深圳检验检疫局克服了种种困难，除从内部增
1268 深圳检验检疫局提请各相关单位、出入境司机注意，《出入境健康检疫申明卡》可在各口岸检验检疫部门
1269 该局同时呼吁相关单位和出入境司机积极配合，共同做好该项工作，切实防止非典经深圳口岸传入传出，但
1270 （国际）美元汇率又跌入低谷。[E]
1271 新华社纽约1月8日电受欧洲中央银行行长讲话的影响，美元汇率在7日经过昙花一现般的回升后8日又

```

The testing command and run:

```

wangyiming@lubo:~/Aaron/CN-EN$ 
wangyiming@lubo:~/Aaron/CN-EN$ 
wangyiming@lubo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train120000/model/moses.ini < ./corpus/test/MT04/LDC2006E43.ch.tok.parsed.mosesxml > ./corpus/test/MT04/LDC2006E43.ch-en.mosesxml.ttt
Defined parameters (per moses.ini or switch):
config: ./train120000/model/moses.ini
cube-pruning-pop-limit: 1000
feature: UnknownWordPenalty WordPenalty PhrasePenalty PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train120000/model/rule-table.gz input-factor=0 output-factor=0 PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train120000/model/glue-grammar input-factor=0 output-factor=0 KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
input-factors: 0
inputtype: 3
mapping: 0 T 0 1 T 1
max-chart-span: 20 1000
non-terminals: X
search-algorithm: 3
weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 TranslationModel0= 0.2 0.2 0.2 0.2 TranslationModel1= 1.0 LM0= 0.5
/smt/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train120000/model/rule-table.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train120000/model/glue-grammar input-factor=0 output-factor=0
FeatureFunction: TranslationModel1 start: 7 end: 7
line=KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LMO start: 8 end: 8
Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading /home/wangyiming/Aaron/Moses/train120000/model/rule-table.gz
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100
*****
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 128x49 | | NUM

Thirdly, test the NIST-MT05 testing data:

```

wangyiming@lubo:~/Aaron/Moses$ 
wangyiming@lubo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train120000/model/moses.ini < ./corpus/test/MT05/LDC2006E38.ch.tok > ./corpus/test/MT05/LDC2006E38.ch-en.ttt
```

It runs as below, the reading of rule-table.gz, etc.:

1:161.64.89.186 - Aaron-20131206 - SSH Secure Shell

File Edit View Window Help

Quick Connect Profiles

```
wangyiming@lobo:~/Aaron/Moses$ wangyiming@lobo:~/Aaron/Moses$ ./smt/mosesdecoder/bin/moses_chart -f ./train120000/model/moses.ini < ./corpus/test/MT05/LDC2006E38.ch.tok > ./corpus/test/MT05/LDC2006E38.ch-en.ttt
Defined parameters (per moses.ini or switch):
config: ./train120000/model/moses.ini
cube-pruning-pop-limit: 1000
feature: UnknownWordPenalty WordPenalty PhrasePenalty PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train120000/model/rule-table.gz input-factor=0 output-factor=0 PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train120000/model/glue-grammar input-factor=0 output-factor=0 KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
input-factors: 0
input-type: 3
mapping: 0 T 0 1 T 1
max-chart-span: 20 1000
non-terminals: X
search-algorithm: 3
weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 TranslationModel0= 0.2 0.2 0.2 0.2 TranslationModel1= 1.0 LMO= 0
.5
/smt/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train120000/model/rule-table.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train120000/model/glue-grammar input-factor=0 output-factor=0
FeatureFunction: TranslationModel1 start: 7 end: 7
line=KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LMO start: 8 end: 8
Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading /home/wangyiming/Aaron/Moses/train120000/model/rule-table.gz
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100
*****
```

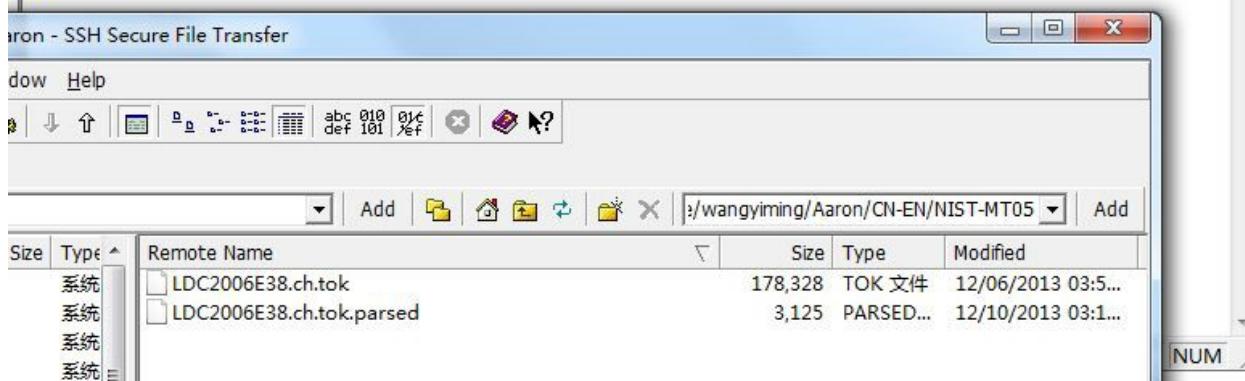
Connected to 161.64.89.186

SSH2 - aes128-cbc - hmac-md5 | 133x37 | NUM

The testing finished as:

Once more, we should first parse the input sentence, then testing:

```
wangyiming@lobo:~/Aaron$ cd CN-EN/
wangyiming@lobo:~/Aaron/CN-EN$ 
wangyiming@lobo:~/Aaron/CN-EN$ java -jar BerkeleyParser-1.7.jar -gr GrammarTrainedCTB7 -inputFile ./NIST-MT05/LDC2006E38.ch.tok -outputFile ./NIST-MT05/LDC2006E38.ch.tok.parsed
```



Testing command:

```
wangyiming@lobo:~/Aaron/Moses$ 
wangyiming@lobo:~/Aaron/Moses$ 
wangyiming@lobo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train120000/model/moses.ini < ./corpus/test/MT05/LDC2006E38.ch.tok.parsed.mosesxml > ./corpus/test/MT05/LDC2006E38.ch-en.mosesxml.ttt
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 120x23 | NUM

Remote Name	Size	Type	Modified	Attribu
LDC2006E38.ch-en.ttt	185,406	T...	12/06/2013 05:40:36 PM	-rw-rw
LDC2006E38.ch-en.Uni.ttt	185,405	T...	12/08/2013 06:09:46 PM	-rw-rw
LDC2006E38.ch.tok	178,328	T...	12/06/2013 03:58:09 PM	-rw-rw
LDC2006E38.ch.tok.parsed.mosesxml	1,856,126	M...	12/10/2013 05:30:58 PM	-rw-rw
LDC2006E38.ch.tok.parsed.UniPhrase.mosesxml	1,879,849	M...	12/10/2013 05:31:35 PM	-rw-rw

Preparing larger corpus-4-Uni-186sever

```
wangyiming@lobo:~/Aaron/Moses$ 
wangyiming@lobo:~/Aaron/Moses$ /smt/mosesdecoder/scripts/training/clean-corpus-n.perl corpus/training-120000/Oxford-dic120000.parsed.
Uni.mosesxml zh en corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.clean 1 80
clean-corpus.perl: processing corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.zh & .en to corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.clean, cutoff 1-80
.....(100000)..
Input sentences: 120000 Output sentences: 68971
wangyiming@lobo:~/Aaron/Moses$
```

2:161.64.89.186 - Aaron-20131206 - SSH Secure File Transfer

Local Name	/	Size	Ty	Remote Name	Size	Type	Modified	Attribu
库			系	Oxford-dic120000.parsed.mosesxml.clean.0.en	2,580,508	EN 文件	12/06/2013 12:1...	-rw-rw-r
Han Lifeng			系	Oxford-dic120000.parsed.mosesxml.clean.0.zh	2,614,666	ZH 文件	12/06/2013 12:1...	-rw-rw-r
计算机			系	Oxford-dic120000.parsed.mosesxml.clean.en	28,663,3...	EN 文件	12/06/2013 11:5...	-rw-rw-r
网络			系	Oxford-dic120000.parsed.mosesxml.clean.zh	31,550,4...	ZH 文件	12/06/2013 11:5...	-rw-rw-r
控制面板			系	Oxford-dic120000.parsed.mosesxml.en	68,078,3...	EN 文件	11/29/2013 06:0...	-rw-rw-r
回收站			系	Oxford-dic120000.parsed.mosesxml.zh	79,160,0...	ZH 文件	11/29/2013 06:0...	-rw-rw-r
控制面板			系	Oxford-dic120000.parsed.Uni.mosesxml.en	69,054,1...	EN 文件	12/02/2013 12:0...	-rw-rw-r
Adobe Reader XI		2,019	快	Oxford-dic120000.parsed.Uni.mosesxml.zh	80,176,4...	ZH 文件	12/02/2013 03:0...	-rw-rw-r

Clean the corpus:

```
wangyiming@lobo:~/Aaron/Moses$ 
wangyiming@lobo:~/Aaron/Moses$ /smt/mosesdecoder/scripts/training/clean-corpus-n.perl corpus/training-120000/Oxford-dic120000.parsed.
Uni.mosesxml zh en corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.clean 1 80
clean-corpus.perl: processing corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.zh & .en to corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.clean, cutoff 1-80
.....(100000)..
Input sentences: 120000 Output sentences: 68971
wangyiming@lobo:~/Aaron/Moses$
```

Training translation model-4-Uni-186sever

Use the following training command:

```
wangyiming@lobo:~/Aaron/Moses$
```

```
/smt/mosesdecoder/scripts/training/train-model.perl -root-dir ./train120000Uni -corpus  
.corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.clean -f zh -e en -alignment grow-  
diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm  
0:3:$HOME/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir /smt/giza-bin-dir/  
&>train.out.Oxford.parse.Uni.zh-en120000
```

```
wangyiming@lobo:~/Aaron/Moses$ /smt/mosesdecoder/scripts/training/train-model.perl -root-dir ./train120000Uni -corpus .corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.clean -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir /smt/giza-bin-dir/ &> train.out.Oxford.parse.Uni.zh-en120000 &
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 80x24

```
top - 12:43:09 up 201 days, 20:09, 2 users, load average: 24.51, 24.14, 24.09  
Tasks: 259 total, 2 running, 257 sleeping, 0 stopped, 0 zombie  
Cpu(s): 3.9%us, 0.0%sy, 96.1%ni, 0.0%id, 0.0%wa, 0.0%hi, 0.0%si, 0.0%st  
Mem: 148540056k total, 34774304k used, 113765752k free, 542288k buffers  
Swap: 50319356k total, 974432k used, 49344924k free, 28795572k cached
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
1100	luyi	39	19	1959m	7672	1212	S	2306	0.0	429352:20	java
17170	wangyimi	20	0	39772	27m	1396	R	93	0.0	0:19.44	mkcls
17164	wangyimi	20	0	17476	1424	960	R	0	0.0	0:00.13	top
1	root	20	0	24340	1556	756	S	0	0.0	0:01.60	init
2	root	20	0	0	0	0	S	0	0.0	0:05.01	kthreadd
3	root	20	0	0	0	0	S	0	0.0	14:43.82	ksoftirqd/0
5	root	20	0	0	0	0	S	0	0.0	6:15.12	kworker/u:0
6	root	RT	0	0	0	0	S	0	0.0	0:18.69	migration/0

Training finished as:

```
30 root      20  0    0    0 S    0  0.0   6:21.05 ksoftirqd/6  
31 root      RT  0    0    0 S    0  0.0   0:44.40 watchdog/6  
[1]+  Done                  /smt/mosesdecoder/scripts/training/train-model.perl -root-dir ./train120000Uni -corpus .corpus/training-120000/Oxford-dic120000.parsed.Uni.mosesxml.clean -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax --target-syntax -lm 0:3:$HOME/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir /smt/giza-bin-dir/ &>train.out.Oxford.parse.Uni.zh-en120000  
wangyiming@lobo:~/Aaron/Moses$
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 80x36

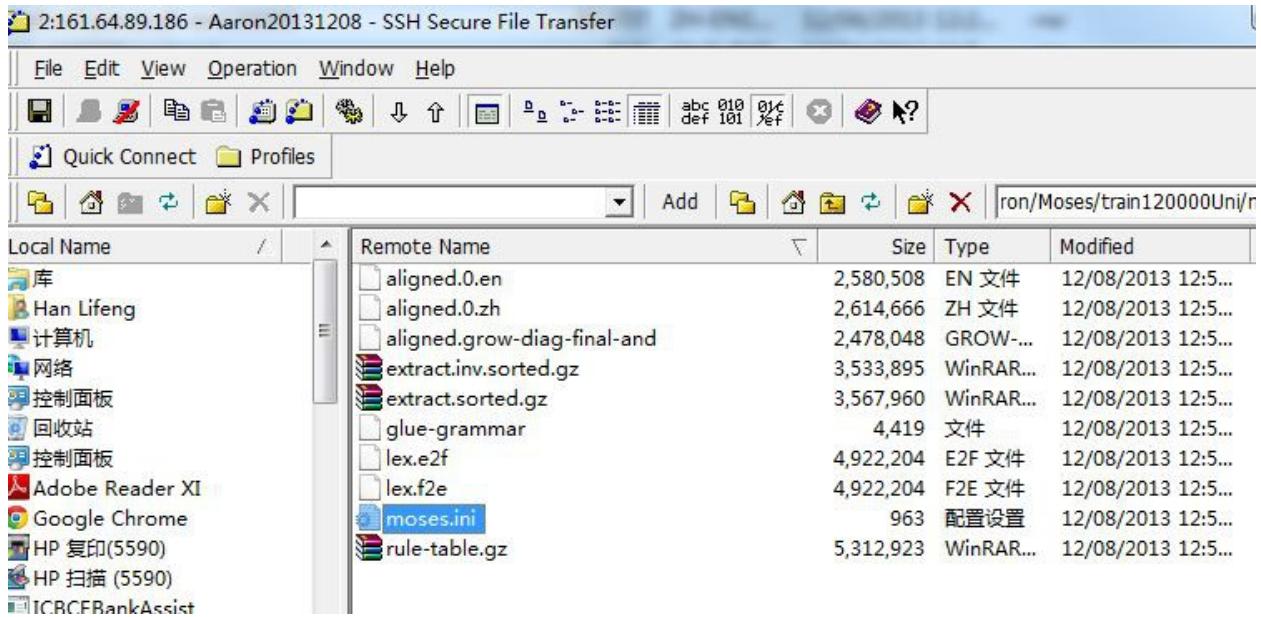
The file-folder will generate:

Aaron20131208 - SSH Secure File Transfer

Operation Window Help						
		Add	Profiles	Remote Name	Size	Type
/	▼			Modified	Attr	
				corpus		Folder
				lm		Folder
				train10000T		Folder
				train120000		Folder
				train12000Uni		Folder
				train.out.Oxford.parse.Uni.zh-en120000	58,146	ZH-EN1...
				train.out.Oxford.parse.zh-en120000	57,557	ZH-EN1...
				tune110001-110000mert.out	939	OUT 文件
				tune110001-110000mert.out.T	145	T 文件

- Aaron20131208 - SSH Secure File Transfer

Operation Window Help						
		Add	Profiles	Remote Name	Size	Type
/	▼			Modified		
				corpus		Folder
				giza.en-zh		Folder
				giza.zh-en		Folder
				model		Folder



Uni120000-Training time from “preparing corpus @ Sun Dec 8 12:42:38 CST 2013” to “create moses.ini @ Sun Dec 8 12:51:57 CST 2013”—total: 9mins+19s

Compare-12000-training time from “(1) preparing corpus @ Fri Dec 6 12:19:27 CST 2013” to “(9) create moses.ini @ Fri Dec 6 12:28:49 CST 2013”—total: 9mins+22s

Testing-4-uni-186sever:

Use the following command for testing the NIST-MT03 testing CN-.EN data

```
wangyiming@lobo:~/Aaron/Moses$ 
wangyiming@lobo:~/Aaron/Moses$ ./smt/mosesdecoder/bin/moses_chart -f ./train12000
0Uni/model/moses.ini < ./corpus/test/LDC2006T04.ch.tok > ./corpus/test/LDC2006T0
4.ch-en.Uni.ttt
```

Testing runs as:

1:161.64.89.186 - Aaron20131208 - SSH Secure Shell

File Edit View Window Help

Quick Connect Profiles

```
Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0 PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0 KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
    input-factors: 0
    inputtype: 3
    mapping: 0 I 0 1 T 1
    max-chart-span: 20 1000
    non-terminals: X
    search-algorithm: 3
    weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 TranslationModel0= 0.2 0.2 0.2 0.2 TranslationModel1= 1.0 LMO= 0.5
/smt/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0
FeatureFunction: TranslationModel1 start: 7 end: 7
line=KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LMO start: 8 end: 8
Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading /home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz
---5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80
---85---90---95---100
*****
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 80x36

The testing translation result is totally the same with the result using the original phrase tags trained model. Why? Is that due to the testing input is plain sentences?

Parse the input sentences, and have a check of the re-testing result. Use the following command to parse the NIST-MT03 testing sentences:

```
wangyiming@lobo:~/Aaron/Moses$ cd ..
wangyiming@lobo:~/Aaron$ cd CN-EN/
wangyiming@lobo:~/Aaron/CN-EN$ wangyiming@lobo:~/Aaron/CN-EN$ java -jar BerkeleyParser-1.7.jar -gr GrammarTrain edCTB7 -inputFile ./LDC2006T04.ch.tok -outputFile ./LDC2006T04.ch.tok.parsed
wangyiming@lobo:~/Aaron/CN-EN$
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 80x24 | |

File - SSH Secure File Transfer

File Help

ze	Type	Remote Name	Size	Type	Modified
系统		BerkeleyParser-1.7.jar	3,092,739	执行文件	10/17/2013 07:3...
系统		GrammarTrainedCTB7	11,602,3...	文件	11/12/2013 03:5...
系统		LDC2006T04.ch.tok	144,371	TOK 文件	12/06/2013 01:1...
系统		LDC2006T04.ch.tok.parsed	412,437	PARSED...	12/10/2013 11:5...
系统		Oxford-dic12000.en	610,057	EN 文件	11/25/2013 11:3...
系统		Oxford-dic12000... ...dic12000... ...dic12000...	1,915,254	PARSED...	11/25/2013 02:0...

Replace the parsed phrase tags into uni-phrase tags and Convert the prased format into moses.xml format:

share-vm > testing > NIST-MT > MT03 > parsed > Uni

新建文件夹

名称	修改日期	类型	大小
berkeleyparsed2mosesxml.pl	10/15/2013 5:24...	PL 文件	1 KB
UniPhrase_LDC2006T04.ch.tok.parsed	12/10/2013 12:0...	PARSED 文件	423 KB
UniPhrase_LDC2006T04.ch.tok.parsed.mosesxml	12/10/2013 12:0...	MOSESXML 文件	1,507 KB

C:\Windows\system32\cmd.exe

```
E:\share-vm\testing\NIST-MT\MT03>
E:\share-vm\testing\NIST-MT\MT03>
E:\share-vm\testing\NIST-MT\MT03>
E:\share-vm\testing\NIST-MT\MT03>"Convert Chinese PENN phrase tags into Universal tags-succeed02.pl"
E:\share-vm\testing\NIST-MT\MT03>cd parsed
E:\share-vm\testing\NIST-MT\MT03\parsed>cd Uni
E:\share-vm\testing\NIST-MT\MT03\parsed\Uni>
E:\share-vm\testing\NIST-MT\MT03\parsed\Uni>
E:\share-vm\testing\NIST-MT\MT03\parsed\Uni>berkeleyparsed2mosesxml.pl < UniPhrase_LDC2006T04.ch.tok.parsed > UniPhrase_LDC2006T04.ch.tok.parsed.mosesxml
E:\share-vm\testing\NIST-MT\MT03\parsed\Uni>
```

Run the translation testing again:

wangyiming@lolo:~/Aaron/Moses\$
wangyiming@lolo:~/Aaron/Moses\$ wangyiming@lolo:~/Aaron/Moses\$./smt/mosesdecoder/bin/moses_chart -f ./train12000Uni/model/moses.ini < ./corpus/test/LDC2006T04.ch.tok.parsed.UniPhrase.mosesxml
> ./corpus/test/LDC2006T04.ch-en.Uni.mosesxml.ttt
Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 80x24

File Explorer view:

Size	Type	Modified
150,339	TTT 文件	12/06/2013
150,339	TTT 文件	12/10/2013
144,371	TOK 文件	12/06/2013
1,542,766	MOSES...	12/10/2013
604,962	EN 文件	12/02/2013
7,416,626	MOSES...	11/29/2013
7,511,416	MOSES...	12/02/2013

It runs as:

1:161.64.89.186 - Profile NameAaron - SSH Secure Shell

File Edit View Window Help

Quick Connect Profiles

```
Model0= 0.2 0.2 0.2 0.2 TranslationModel1= 1.0 LM0= 0.5
/smt/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4
path=/home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0
FeatureFunction: TranslationModel1 start: 7 end: 7
line=KENLM lazyken=0 name=LM0 factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LM0 start: 8 end: 8
Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading /home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80
---85---90---95---100
*****
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 80x24

File Explorer view:

Size	Type	Modified
150,339	TTT 文件	12/06/2013
150,339	TTT 文件	12/10/2013
144,371	TOK 文件	12/06/2013
1,542,766	MOSES...	12/10/2013

The translation result is different with the translation that uses the plain input text.

Secondly, we used the NIST-MT04 CN-EN test data:

```
wangyiming@lobo:~/Aaron/Moses$ wangyiming@lobo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train12000Uni/model/moses.ini < ./corpus/test/MT04/LDC2006E43.ch.tok > ./corpus/test/MT04/LDC2006E43.ch-en.Uni.ttt
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 80x40

Testing once more using the parsed corpus:

```
wangyiming@lobo:~/Aaron/Moses$ wangyiming@lobo:~/Aaron/Moses$ wangyiming@lobo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train12000Uni/model/moses.ini < ./corpus/test/MT04/LDC2006E43.ch.tok.parsed.UniPhrase.mosesxml > ./corpus/test/MT04/LDC2006E43.ch-en.Uniphrase.mosesxml.ttt
```

Defined parameters (per moses.ini or switch):
config: ./train12000Uni/model/moses.ini
cube-pruning-pop-limit: 1000
feature: UnknownWordPenalty WordPenalty PhrasePenalty PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0 PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0 KENLM lazyken=0 name=LM0 factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
input-factors: 0
inputputype: 3
mapping: 0 T 0 1 T 1
max-chart-span: 20 1000
non-terminals: X
search-algorithm: 3
weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 TranslationModel0= 0.2 0.2 0.2 0.2 TranslationModel1= 1.0 LM0= 0.5
/smt/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0
FeatureFunction: TranslationModel1 start: 7 end: 7
line=KENLM lazyken=0 name=LM0 factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LM0 start: 8 end: 8
Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading /home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 120x54

It finished as:

Thirdly, test the NIST-MT05 testing data:

```
wangyiming@lolo:~/Aaron/Moses$ /smt/mosesdecoder/bin/moses_chart -f ./train120000Uni/mo  
del/moses.ini < ./corpus/test/MT05/LDC2006E38.ch.tok > ./corpus/test/MT05/LDC2006E38.ch  
-en.Uni.ttt
```

The testing runs as:

```

-----
wangyiming@lobo:~/Aaron/Moses$ wangyiming@lobo:~/Aaron/Moses$ ./smt/mosesdecoder/bin/moses_chart -f ./train12000Uni/model/moses.ini < ./corpus/test/MT05/LDC2006E38.ch.tok > ./corpus/test/MT05/LDC2006E38.ch-en.Uni.ttt
Defined parameters (per moses.ini or switch):
config: ./train12000Uni/model/moses.ini
cube-pruning-pop-limit: 1000
feature: UnknownWordPenalty WordPenalty PhrasePenalty PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0 PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0 KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
input-factors: 0
inputputtype: 3
mapping: 0 T 0 1 T 1
max-chart-span: 20 1000
non-terminals: X
search-algorithm: 3
weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 TranslationModel0= 0.2 0.2 0.2 0.2 TranslationModel1= 1.0 LMO= 0.5
/smt/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0
FeatureFunction: TranslationModel1 start: 7 end: 7
line=KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LMO start: 8 end: 8
Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading /home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---
-90---95---100
*****

```

Connected to 161.64.89.186

SSH2 - aes128-cbc - hmac-md5 87x46



NU

Testing once more using the parsed corpus:

```
wangyiming@lobo:~/Aaron/Moses$ ./smt/mosesdecoder/bin/moses_chart -f ./train12000Uni/model/moses.ini < ./corpus/test/MT05/LDC2006E38.ch.tok.parsed.UniPhrase.mosesxml > ./corpus/test/MT05/LDC2006E38.ch-en.Uniphrase.mosesxml.ttt
Defined parameters (per moses.ini or switch):
config: ./train12000Uni/model/moses.ini
cube-pruning-pop-limit: 1000
feature: UnknownWordPenalty WordPenalty PhrasePenalty PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0 PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0 KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
input-factors: 0
inputtype: 3
mapping: 0 T 0 1 T 1
max-chart-span: 20 1000
non-terminals: X
search-algorithm: 3
weight: UnknownWordPenalty0= 1 WordPenalty0= -1 PhrasePenalty0= 0.2 TranslationModel0= 0.2 0.2 0.2 0.2 TranslationModel1= 1.0 LMO= 0.5
/smt/mosesdecoder/bin
line=UnknownWordPenalty
FeatureFunction: UnknownWordPenalty0 start: 0 end: 0
line=WordPenalty
FeatureFunction: WordPenalty0 start: 1 end: 1
line=PhrasePenalty
FeatureFunction: PhrasePenalty0 start: 2 end: 2
line=PhraseDictionaryMemory name=TranslationModel0 table-limit=20 num-features=4 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz input-factor=0 output-factor=0
FeatureFunction: TranslationModel0 start: 3 end: 6
line=PhraseDictionaryMemory name=TranslationModel1 num-features=1 path=/home/wangyiming/Aaron/Moses/train12000Uni/model/glue-grammar input-factor=0 output-factor=0
FeatureFunction: TranslationModel1 start: 7 end: 7
line=KENLM lazyken=0 name=LMO factor=0 path=/home/wangyiming/Aaron/Moses/lm/news-commentary-v8.fr-en.blm.en order=3
FeatureFunction: LMO start: 8 end: 8
Start loading text SCFG phrase table. Moses format : [0.000] seconds
Reading /home/wangyiming/Aaron/Moses/train12000Uni/model/rule-table.gz
---5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95---100
*****
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 120x54 | NUM

Build CN-EN tree-to-string translation using moses:

Training translation Corpus: source-parsed, target-parsed

Training:

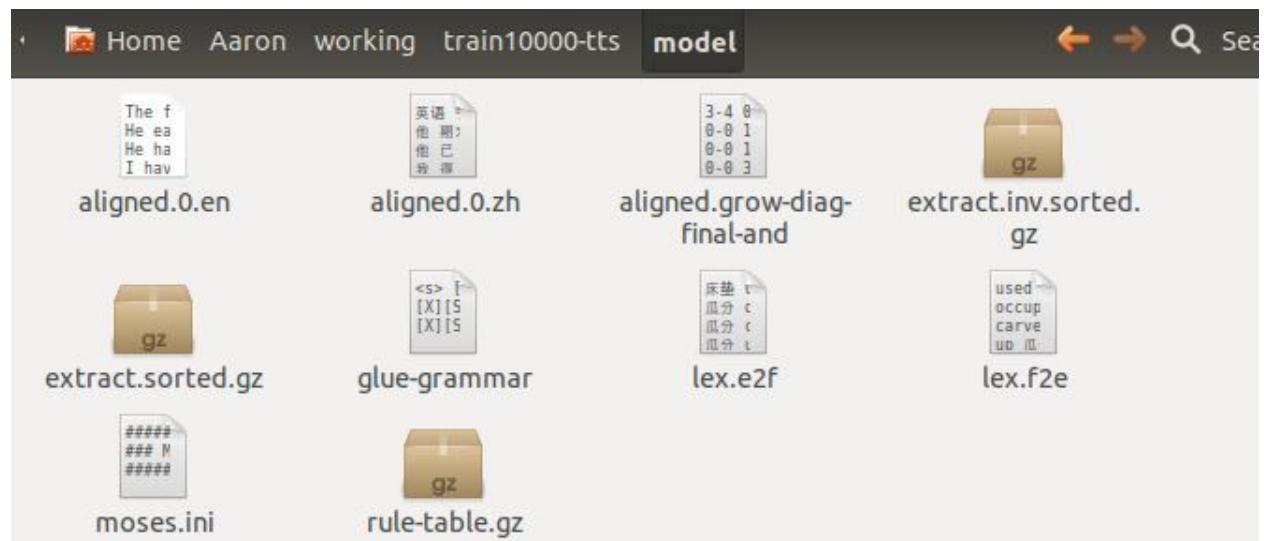
```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts -corpus ~/Aaron/corpus/training/train100000xf/Oxford-dic10000.parsed.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.zh-en10000-tts &
```

```

nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts -corpus ~/Aaron/corpus/training/train10000xf/Oxford-dic10000.parsed.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& training.out.Oxford.parse.zh-en10000-tts &
[1] 2354
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ jobs
[1]+  Running                  ~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts -corpus ~/Aaron/corpus/training/train10000xf/Oxford-dic10000.parsed.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools &>training.out.Oxford.parse.zh-en10000-tts &
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
```

```

 11 root      RT   0      0      0 S    0  0.0  0:00.02 watchdog/1
 12 root      0 -20     0      0 S    0  0.0  0:00.00 cpuset
[1]+  Done                  ~Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts -corpus ~/Aaron/corpus/training/train10000xf/Oxford-dic10000.parsed.mosesxml -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools &>training.out.Oxford.parse.zh-en10000-tts
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ 
```



Testing:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_chart -f train10000-tts/model/moses.ini < ~/Aaron/corpus/test/tree-to-string/Oxford-dic11001-12000.zh.parsed.mosesxml > Oxford-dic11001-12000.zh.parsed.tts-out
```

```
      1   0   0   0   0  
      1   0   0  
      4   0  
      1  
BEST TRANSLATION: 200 S -> S </s> : 0-0 : c=-0.460 core=(0.000,-1.000,1.000,0.00  
0,0.000,0.000,0.000,0.000,0.000) [0..18] 196 [total=-316.465] core=(-300.000,-1  
9.000,36.000,-70.037,0.000,-32.019,0.000,16.998,-78.504)  
Translation took 0.020 seconds  
End. : [22.000] seconds  
Name:moses_chart          VmPeak:319520 kB           VmRSS:161260 kB RSSMax:162652 kB  
user:2.412      sys:9.941      CPU:12.353      real:22.394  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

```
commands.txt ✘ Oxford-dic11001-12000.zh.parsed.tts-out ✘  
</tree> label="TOP"> label="S"> label="JJ"> Hit label="SINV"> </tree> label=","> </tree>  
<tree label="TOP"> 牝依 基督教 </tree>  
十字架 label="TOP"> 基督教 <tree 象征 </tree>  
<tree label="VBP"> label="NN"> intruder 异教徒 宣传 基督教 </tree>  
<tree label="S"> 改 label="VP"> 基督教 <tree </tree>  
基督教 义 label="S"> <tree 三位一体 指 <tree label="TOP"> 圣父 tasteless 圣子 colorless 圣灵 </tree>  
十字架 label="TOP"> 基督教 <tree label="PRP"> 徵 </tree>  
label="IN"> label="TOP"> 克里斯蒂娜 <tree <tree <tree 继任 label="TOP"> </tree>  
恭祝 圣诞 <tree Ramsay 贺新福 label="FRAG">  
label="FRAG"> <tree label="NNP"> 互 <tree 贺卡 colorless label="PP"> </tree>  
label="WRB"> <tree label="NNP"> label="S"> <tree label="TOP"> label="TOP"> <tree label="NP"> </tree>  
they </tree> <tree label="NNP"> label="TOP"> <tree label="S"> <tree </tree>  
<tree label="TOP"> <tree label="TOP"> <tree label="NNP"> </tree> label="WHNP">  
label="NNP"> label="WHNP"> 停 label="S"> <tree </tree>  
label="NNP"> sergeant label="S"> 火鸡 label="TOP"> label="S"> <tree <tree label="NP"> </tree>
```

Build CN-EN tree-to-string translation using moses-2:

Training translation corpus: source-parsed, target-plain

Training:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts-2 -corpus ~/Aaron/corpus/training/train10000xf/tts/cn-en/Oxford-dic10000-tts-cn-en -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& translation-Oxford-tts-zh-en10000 &
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts-2 -corpus ~/Aaron/corpus/training/train10000xf/tts/cn-en/Oxford-dic10000-tts-cn-en -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& translation-Oxford-tts-zh-en10000 &
[1] 3684
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ jobs
[1]+  Running                  ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts-2 -corpus ~/Aaron/corpus/training/train10000xf/tts/cn-en/Oxford-dic10000-tts-cn-en -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools &>translation-Oxford-tts-zh-en10000 &
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

```
10 root      20    0    0    0    0 S    0  0.0  0:05.07 ksoftirqd/1
11 root      RT    0    0    0 S    0  0.0  0:01.37 watchdog/1
12 root      0 -20    0    0 S    0  0.0  0:00.00 cpuset
[1]+  Done                  ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts-2 -corpus ~/Aaron/corpus/training/train10000xf/tts/cn-en/Oxford-dic10000-tts-cn-en -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools &>translation-Oxford-tts-zh-en10000
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Testing:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_c  
hart -f train10000-tts-2/model/moses.ini < ~/Aaron/corpus/test/tree-to-string/Ox  
ford-dic11001-12000.zh.parsed.mosesxml > Oxford-dic11001-12000.zh.parsed.tts-out  
-2
```

```
1 0 0 0 0 0  
5 0 0 0 0 0  
1 0 0 0 0 0  
1 0 0 0 0 0  
15 0 0 0 0 0  
1  
BEST TRANSLATION: 954 S -> S </s> :0-0 : c=-0.460 core=(0.000,-1.000,1.000,0.00  
0,0.000,0.000,0.000,0.000,0.000) [0..18] 937 [total=-323.282] core=(-300.000,-1  
8.000,34.000,-19.144,-28.310,-12.946,-17.314,15.998,-97.074)  
Translation took 0.020 seconds  
End. : [25.000] seconds  
Name:moses_chart VmPeak:265848 kB VmRSS:121316 kB RSSMax:123312 kB  
user:4.048 sys:10.593 CPU:14.641 real:24.648  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Oxford-dic11001-12000.zh.parsed.tts-out-2 (~/Aaron/working) - gedit

commands.txt Oxford-dic11001-12000.zh.parsed.tts-out Oxford-dic11001-12000.zh.parsed.tts-out-2

Roman is that the world on the great city .
she has 坂依 基督教 .
十字架 's 基督教 of 象征 .
he to foreign bows 异教徒 program 基督教 .
he recently 改 letter 基督教 the .
基督教 义 welding of 三位一体 a of 's 圣父 , 圣子 and 圣灵 .
十字架 's 基督教 of as 徵 .
the 've to 克里斯蒂娜 a his 继任 people .
恭祝 圣诞 , and 贺新禧 !
people in the Christmas 互 made amends for his rude remarks by giving 贺卡 and present .
of the past few Christmas holiday 've bless quirk 平静 .
this year of Christmas is on Monday at .
the country of people 've Christmas you ?
Christmas train 停 bound for .
Christmas billows eat 火鸡 is in England of traditional .
Christmas 前夕 , Mr. Smith pleased , because he received friends of many by .
Christmas 节期 圣诞 节假日 from 十二月 二十四日 of 圣诞 straps a day , to 一月 五日 of 显灵 节前 a day
we are to church honour 圣诞前夕 .
筹办 圣诞 party must activates I 智 poor realised 竭 .
they in 圣诞夜 will do whatever it ?
we buy the capacitor little value to 点缀 圣诞树 .

Evaluation:

Using the following command, we calculate the BLEU score of the 1000 translated sentences:

Export result and evaluate using Asiya-tool?

Build CN-EN tree-to-string translation using moses-Uni:

Training translation corpus: source-parsed, target-plain

Training:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-tts-Uni -corpus ~/Aaron/corpus/training/train10000xf/tts/cn-en/Oxford-dic10000-tts-cn-en-Uni -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& translation-Oxford-tts-Uni-zh-en10000 &
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
3924	nlp2ct	20	0	24460	12m	1460	R	100	0.2	0:14.18	mkcls
940	root	20	0	388m	151m	12m	S	1	2.3	8:43.27	Xorg
1660	nlp2ct	20	0	1322m	84m	34m	S	1	1.3	11:31.69	compiz
2222	nlp2ct	20	0	662m	43m	20m	S	1	0.7	0:15.62	gedit
1	root	20	0	24436	2360	1352	S	0	0.0	0:01.15	init
2	root	20	0	0	0	0	S	0	0.0	0:00.06	kthreadd
3	root	20	0	0	0	0	S	0	0.0	0:06.80	ksoftirqd/0
5	root	20	0	0	0	0	S	0	0.0	0:00.89	kworker/u:0

```
17 root      20  0    0    0 S    0  0.0  0:01.57 sync_supers  
[1]+ Done                  ~/Aaron/Moses/mosesdecoder/scripts/training/train-  
model.perl -root-dir train10000-tts-Uni -corpus ~/Aaron/corpus/training/train10000xf/tts/cn-en/Oxford-dic10000-tts-cn-en-Uni -f zh -e en -alignment grow-diag-f  
inal-and -hierarchical -glue-grammar --source-syntax -lm 0:3:$HOME/Aaron/lm/news  
-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools  
&>translation-Oxford-tts-Uni-zh-en10000  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Testing:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_c  
hart -f train10000-tts-Uni/model/moses.ini < ~/Aaron/corpus/test/tree-to-string/  
Oxford-dic11001-12000.zh.parsed.Uni.mosesxml > Oxford-dic11001-12000.zh.parsed.t  
ts-out-Uni
```

```
      5   0   0   0   0  
      1   0   0   0  
      1   0   0  
     15   0  
      1  
BEST TRANSLATION: 1143 S -> S </s> :0-0 : c=-0.460 core=(0.000,-1.000,1.000,0.0  
00,0.000,0.000,0.000,0.000) [0..18] 1122 [total=-323.282] core=(-300.000,  
-18.000,34.000,-19.145,-28.310,-12.946,-17.314,15.998,-97.074)  
Translation took 0.030 seconds  
End. : [26.000] seconds  
Name:moses_chart           VmPeak:265568 kB          VmRSS:115248 kB RSSMax:122844 kB  
user:4.564      sys:11.157      CPU:15.721      real:26.023  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

```
Oxford-dic11001-120...h.parsed.tts-out-Uni ✘  
Roman is that the world on the great city .  
she has 坂依 基督教 .  
十字架 's 基督教 of 象征 .  
he to foreign bows 异教徒 program 基督教 .  
he recently 改 letter 基督教 the .  
基督教 义 welding of 三位一体 a of 's 圣父 , 圣子 and 圣灵 .  
十字架 's 基督教 of as 徵 .  
the 've to 克里斯蒂娜 a his 继任 people .  
恭祝 圣诞 , and 贺新福 !  
people in the Christmas 互 made amends for his rude remarks by giving 贺卡 and present .  
of the past few Christmas holiday 've bless quirk 平静 .  
this year of Christmas is on Monday at .  
the country of people 've Christmas you ?  
Christmas train 停 bound for .  
Christmas billows eat 火鸡 is in England of traditional .  
Christmas 前夕 , Mr. Smith pleased , because he received friends of many by .  
Christmas 节期 圣诞 节假日 from 十二月二十四日 of 圣诞 straps a day , to 一月五日 of 显灵 节前 a day  
we are to church honour 圣诞前夕 .  
筹办 圣诞 party must activates I 智 poor realised 竭 .  
they in 圣诞夜 will do whatever it ? . . .
```

Tree-to-string: Using the ori-phrase and uni-phrase, the evaluation scores using Asiya-tool are different.

Build CN-EN string-to-tree translation using moses-ori:

Training translation Corpus: source-plain, target-parsed.

Training:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-stt-ori -corpus ~/Aaron/corpus/training/train10000xf/stt/cn-en/Oxford-dic10000.stt-cn-en-ori -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& translation-Oxford-stt-ori-zh-en10000 &
```

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-stt-ori -corpus ~/Aaron/corpus/training/train10000xf/stt/cn-en/Oxford-dic10000.stt-cn-en-ori -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& translation-Oxford-stt-ori-zh-en10000 &
[1] 4124
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ jobs
[1]+  Running                  ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-stt-ori -corpus ~/Aaron/corpus/training/train10000xf/stt/cn-en/Oxford-dic10000.stt-cn-en-ori -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools &>translation-Oxford-stt-ori-zh-en10000 &
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

```
 10 root      20  0    0    0    0 S    0  0.0  0:03.53 ksorttrqd/1
 11 root      RT  0    0    0 S    0  0.0  0:01.54 watchdog/1
[1]+  Done                  ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-stt-ori -corpus ~/Aaron/corpus/training/train10000xf/stt/cn-en/Oxford-dic10000.stt-cn-en-ori -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools &>translation-Oxford-stt-ori-zh-en10000
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Testing:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/bin/moses_chart -f train10000-stt-ori/model/moses.ini < ~/Aaron/corpus/test/string-to-tree/Oxford-dic11001-12000.zh > Oxford-dic11001-12000.zh-en.stt.out-ori
```

```

      1   0   0   0
      1   0   0
      7   0
      1
BEST TRANSLATION: 2473 Q -> Q </s> :0-0 : c=-0.460 core=(0.000,-1.000,1.000,0.0
00,0.000,0.000,0.000,0.000) [0..18] 2471 [total=-324.858] core=(-300.000,
-18.000,34.000,-26.661,-28.979,-8.618,-16.301,15.998,-99.089)
Translation took 0.020 seconds
End. : [25.000] seconds
Name:moses_chart          VmPeak:297948 kB           VmRSS:158348 kB RSSMax:163592 kB
user:5.132      sys:10.109     CPU:15.241           real:24.641
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Oxford-dic11001-12000.zh-en.stt.out-ori (~/Aaron/working) - gedit

Open Save Undo Redo Cut Copy Paste Find Replace

OXford-dic11001-12000.zh-en.stt.out-ori

Roman is that the world on the great city .
her hydraulics 版依 基督教 .
十字架 is 基督教 the badge of maturity .
his abroad bows 异教徒 宣传 基督教 .
he recently 改 letter 基督教 .
基督教 义 welding of 三位一体 a of is 圣父 , 圣子 and 圣灵 .
十字架 is 基督教 of genius .
the hydraulics to 克里斯蒂娜 make his 继任 people .
恭祝 圣诞 , and 贺新禧 !
people in Christmas 互 staff 贺卡 and present .
to of robbers Christmas holiday been very at peace .
this year of Christmas is 星期一 .
the country 's people across Christmas ?
Christmas train 停 bound for a debate
Christmas drunkard eat 火鸡 is in England of traditional .
Christmas 前夕 , Mr. Smith was glad , because he received friends of many letters .
Christmas 节期 圣诞 节假日 from 十二月二十四日 of 圣诞 ago , to 一月五日 of 显灵 节前 a Darkness
we church 庆祝 圣诞 前夕 .
筹办 圣诞 party must antedate I 智 poor realised 竭 .
they in 圣诞 夜 will do anything about it ?
we buy the capacitor little value to 点缀 圣诞树 .

Build CN-EN string-to-tree translation using moses-uni:

Training translation Corpus: source-plain, target-parsed.

Training:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~ /Aaron/Moses/mosesdecoder/scripts/training/train-model.perl -root-dir train10000-stt-uni -corpus ~/Aaron/corpus/training/train10000xf/stt/cn-en/Oxford-dic10000.stt-cn-en-uni -f zh -e en -alignment grow-diag-final-and -hierarchical -glue-grammar --target-syntax -lm 0:3:$HOME/Aaron/lm/news-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools >& translation-Oxford-stt-uni-zh-en10000 &
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
4401	nlp2ct	20	0	37616	24m	1792	R	100	0.4	0:05.43	GIZA++
940	root	20	0	393m	156m	12m	S	1	2.4	9:34.05	Xorg
1660	nlp2ct	20	0	1322m	84m	34m	S	1	1.3	12:12.47	compiz
1622	nlp2ct	20	0	109m	1576	1024	S	0	0.0	4:20.59	VBoxClient
1682	nlp2ct	20	0	1461m	59m	21m	S	0	0.9	1:19.87	nautilus
1968	nlp2ct	20	0	530m	25m	16m	S	0	0.4	1:14.63	gnome-terminal

```
15 root      0 -20      0      0 S      0  0.0  0:00.00 netns
[1]+  Done                  ~ /Aaron/Moses/mosesdecoder/scripts/training/train-
model.perl -root-dir train10000-stt-uni -corpus ~/Aaron/corpus/training/train10000xf/stt/cn-en/Oxford-dic10000.stt-cn-en-uni -f zh -e en -alignment grow-diag-f
inal-and -hierarchical -glue-grammar --target-syntax -lm 0:3:$HOME/Aaron/lm/news
-commentary-v8.fr-en.blm.en:8 -external-bin-dir ~/Aaron/Moses/mosesdecoder/tools
&>translation-Oxford-stt-uni-zh-en10000
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Testing:

```
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$ ~/Aaron/Moses/mosesdecoder/bin/moses_c  
hart -f train10000-stt-uni/model/moses.ini < ~/Aaron/corpus/test/string-to-tree/  
Oxford-dic11001-12000.zh > Oxford-dic11001-12000.zh-en.stt.out-uni
```

```
1   0   0   0  
1   0   0  
7   0  
1  
BEST TRANSLATION: 2469 Q -> Q </s> :0-0 : c=-0.460 core=(0.000,-1.000,1.000,0.0  
00,0.000,0.000,0.000,0.000,0.000) [0..18] 2467 [total=-324.858] core=(-300.000,  
-18.000,34.000,-26.661,-28.979,-8.618,-16.301,15.998,-99.089)  
Translation took 0.030 seconds  
End. : [25.000] seconds  
Name:moses_chart      VmPeak:297124 kB      VmRSS:157604 kB RSSMax:163204 kB  
user:5.412      sys:10.017     CPU:15.429      real:24.695  
nlp2ct@nlp2ct-VirtualBox:~/Aaron/working$
```

Oxford-dic11001-12000.zh-en.stt.out-uni (~/Aaron/working) - gedit

Open Save Undo Redo Cut Copy Paste Find Replace

Roman is that the world on the great city .
her hydraulics 叢依 基督教 .
十字架 is 基督教 the badge of maturity .
his abroad bows 异教徒 宣传 基督教 .
he recently 改 letter 基督教 .
基督教 义 welding of 三位一体 a of is 圣父 , 圣子 and 圣灵 .
十字架 is 基督教 of genius .
the hydraulics to 克里斯蒂娜 make his 继任 people .
恭祝 圣诞 , and 贺新禧 !
people in Christmas 互 staff 贺卡 and present .
to of robbers Christmas holiday been very at peace .
this year of Christmas is 星期一 .
the country 's people across Christmas ?
Christmas train 停 bound for a debate
Christmas drunkard eat 火鸡 is in England of traditional .
Christmas 前夕 , Mr. Smith was glad , because he received friends of many letters .
Christmas 节期 圣诞 节假日 from 十二月 二十四日 of 圣诞 ago , to 一月 五日 of 显灵 节前 a Darkness
we church 庆祝 圣诞 前夕 .
筹办 圣诞 party must antedate I 智 poor realised 竭 .
they in 圣诞 夜 will do anything about it ?
we buy the capacitor little value to 点缀 圣诞树 .
Somebody the the switch , 圣诞树 on all 灯 expenses 亮 the knotted .

String-to-tree: Using the ori-phrase and uni-phrase, the evaluation scores using Asiya-tool are the same, why?

Build PT->CN tree-to-tree translation model with ori-phrase tags on 186-server:

Prepare corpus:

Train language model:

Train language model of CN uses the 120,000 Oxford simplified CN sentences.

Use the following command to check the moses installation on the server:

1:161.64.89.186 - Aaron20131212 - SSH Secure Shell

File Edit View Window Help

Quick Connect Profiles

```
wangyiming@lobo:~$ ls
Aaron corpus CWMT2013 download james lm mert-work moses process work
wangyiming@lobo:~$ cd lm/
wangyiming@lobo:~/lm$ ls
chinese-nlpirc.blm
wangyiming@lobo:~/lm$ ls
chinese-nlpirc.blm
wangyiming@lobo:~/lm$ cd ..
wangyiming@lobo:~/~ cd moses/
wangyiming@lobo:~/moses$ ls
boost boost_1_52_0 giza-pp irstlm-5.80.03 mosesdecoder
wangyiming@lobo:~/moses$ 
wangyiming@lobo:~/moses$ ls
wangyiming@lobo:~/moses$ cd irstlm-5.80.03/
wangyiming@lobo:~/moses/irstlm-5.80.03$ ls
aclocal.m4 config.h.in~ Copyright libtool missing stamp-h1
autom4te.cache config.log depcomp ltmain.sh README
bin config.status include m4 regenerate-makefiles.sh
config.guess config.sub install-sh Makefile RELEASE
config.h configure irstlm Makefile.am scripts
config.h.in configure.in lib Makefile.in src
wangyiming@lobo:~/moses/irstlm-5.80.03$ cd scripts/
wangyiming@lobo:~/moses/irstlm-5.80.03/scripts$ ls
add-start-end.sh goograms2ngrams.pl Makefile.in rm-start-end.sh wrapper
build-lm-qsub.sh lm-stat.pl mdtsel.sh sort-lm.pl
build-lm.sh Makefile merge-sublm.pl split-dict.pl
build-sublm.pl Makefile.am ngram-split.pl split-ngt.sh
wangyiming@lobo:~/moses/irstlm-5.80.03/scripts$
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 95x29

Put the 120000 sentences chinese corpus:

Add	Remote Name	Size	Type	Modified
	Oxford-dictionary.zh.tok	6,667,108	TOK 文件	03/19/

Oxford-dictionary.zh

```
1 英语 字母表 中 的 第 一 个 字 母 是 A 。 LF
2 他 期 末 考 试 四 门 功 课 得 优 。 LF
3 他 已 在 洛 杰 斯 找 到 一 份 工 作 。 LF
4 我 得 了 重 感 冒 ， 总 流 鼻 涕 。 LF
5 他 在 外 面 淋 了 一 天 雨 ， 因 此 患 了 重 感 冒 。 LF
6 我 们 只 得 对 此 不 再 抱 有 希 望 。 LF
7 他 是 那 么 会 晕 船 ， 所 以 他 总 搭 飞 机 。 LF
8 由 邮 递 员 同 时 分 送 的 一 批 邮 件 。 LF
9 我 的 私 人 信 件 和 一 批 通 知 混 在 一 起 了 。 LF
10 会 计 签 发 的 一 批 支 票 。 LF
```

Use the following command to add start and end symbol “<s> </s>”:

```
wangyiming@lobo:~$ /smt/irstlm-5.80.03/scripts/add-start-end.sh < Aaron/Moses/lm/CN/Oxford-dictionary.zh.tok > Aaron/Moses/lm/CN/Oxford-dictionary.zh.sb
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 95x29 | NUM

Use the following command for the generation of language model:

```
wangyiming@lobo:~$ export IRSTLM=$Home/smt/irstlm-5.80.03; /smt/irstlm-5.80.03/scripts/build-lm.sh -i Aaron/Moses/lm/CN/Oxford-dictionary.zh.sb -t Aaron/Moses/lm/CN/tmp -p -s improved-kneser-ney -o Aaron/Moses/lm/CN/Oxford-dictionary.zh.lm
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 95x29 | NUM

```
wangyiming@lobo:~$ export IRSTLM=$Home/smt/irstlm-5.80.03; /smt/irstlm-5.80.03/scripts/build-lm.sh -i Aaron/Moses/lm/CN/Oxford-dictionary.zh.sb -t Aaron/Moses/lm/CN/tmp -p -s improved-kneser-ney -o Aaron/Moses/lm/CN/Oxford-dictionary.zh.lm
Temporary directory Aaron/Moses/lm/CN/tmp does not exist
creating Aaron/Moses/lm/CN/tmp
Extracting dictionary from training corpus
Splitting dictionary into 3 lists
Extracting n-gram statistics for each word list
Important: dictionary must be ordered according to order of appearance of words in data used to generate n-gram blocks, so that sub language model blocks results ordered too
dict.000
dict.001
dict.002
$bin/ngt -i="$inpfle" -n=$order -gootout=y -o="$gzip -c > $tmpdir/ngram.${sdict}.gz" -fd="$tmpdir/$sdict" $dictionary -iknstat="$tmpdir/ikn.stat.$sdict" >> $logfile 2>&1
Estimating language models for each word list
dict.000
dict.001
dict.002
$scr/build-sublm.pl $verbose $prune $smoothing "cat $tmpdir/ikn.stat.dict.*" --size $order --ng rams "$gunzip -c $tmpdir/ngram.${sdict}.gz" -sublm $tmpdir/lm.$sdict >> $logfile 2>&1
Merging language models into Aaron/Moses/lm/CN/Oxford-dictionary.zh.lm
Cleaning temporary directory Aaron/Moses/lm/CN/tmp
Removing temporary directory Aaron/Moses/lm/CN/tmp
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 | 95x29 | NUM

The generated language model:

Remote Name	Size	Type	Modified	Attributes
Oxford-dictionary.zh.lm.gz	4,701,426	WinRAR...	12/12/2013 08:1...	-rw-rw-r--
Oxford-dictionary.zh.sb	7,747,108	SB 文件	12/12/2013 08:0...	-rw-rw-r--
Oxford-dictionary.zh.tok	6,667,108	TOK 文件	03/19/2013 01:3...	-rw-rw-r--

Use the following command to compile the language model, generate the “”:

```
wangyiming@lobo:~$ /smt/irstlm-5.80.03/src/compile-lm -text Aaron/Moses/lm/CN/Oxford-dictionary
.zh.lm.gz Aaron/Moses/lm/CN/Oxford-dictionary.zh.arpa
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 95x29 NUM

```
wangyiming@lobo:~$ /smt/irstlm-5.80.03/src/compile-lm -text Aaron/Moses/lm/CN/Oxford-dictionary
.zh.lm.gz Aaron/Moses/lm/CN/Oxford-dictionary.zh.arpa
inpfile: Aaron/Moses/lm/CN/Oxford-dictionary.zh.lm.gz
outfile: Aaron/Moses/lm/CN/Oxford-dictionary.zh.arpa
loading up to the LM level 1000 (if any)
dub: 10000000
Language Model Type of Aaron/Moses/lm/CN/Oxford-dictionary.zh.lm.gz is 1
Language Model Type is 1
iARPA
loadtxt_ram()
1-grams: reading 37992 entries
done level 1
2-grams: reading 361419 entries
done level 2
3-grams: reading 209786 entries
done level 3
done
OOV code is 37991
OOV code is 37991
Saving in txt format to Aaron/Moses/lm/CN/Oxford-dictionary.zh.arpa
savetxt: Aaron/Moses/lm/CN/Oxford-dictionary.zh.arpa
save: 37992 1-grams
save: 361419 2-grams
save: 209786 3-grams
done
wangyiming@lobo:~$
```

Connected to 161.64.89.186 SSH2 - aes128-cbc - hmac-md5 95x29 NUM

Remote Name	Size	Type	Modified	Attributes
Oxford-dictionary.zh.arpa	15,364,649	ARPA 文...	12/12/2013 08:2...	-rw-rw-r--
Oxford-dictionary.zh.lm.gz	4,701,426	WinRAR...	12/12/2013 08:1...	-rw-rw-r--
Oxford-dictionary.zh.sb	7,747,108	SB 文件	12/12/2013 08:0...	-rw-rw-r--
Oxford-dictionary.zh.tok	6,667,108	TOK 文件	03/19/2013 01:3...	-rw-rw-r--

The content of the zh.arpa file:

Oxford-dictionary.zh Oxford-dictionary.zh.arpa

```
1 LF
2 \data\LF
3 ngram 1= 37992LF
4 ngram 2= 361419LF
5 ngram 3= 209786LF
6 LF
7 LF
8 \1-grams:LF
9 -5.84994 →<s>→-1.00137LF
10 -3.55218 →英语→-0.628051LF
11 -5.03702 →字母表→-0.694482LF
12 -2.51118 →中→-0.692795LF
13 -1.28505 →的→-0.839764LF
14 -3.5863 →第一→-0.70014LF
15 -2.46891 →个→-0.550793LF
16 -4.17784 →字母→-0.579907LF
17 -1.93236 →是→-0.741366LF
18 -5.00484 →A→-0.214711LF
19 -1.1442 →。→-2.67583LF
20 -1.07178 →</s>LF
89023 -4.32159 →我们深入LF
89024 -4.31424 →我们雇LF
89025 -3.35469 →我们介绍→-0.134097LF
89026 -3.40422 →我们因为→-0.567763LF
89027 -4.12123 →我们具有LF
89028 -4.2685 →我们计算LF
89029 -4.28012 →我们有所LF
89030 -2.39817 →我们听到→-0.132928LF
89031 -4.25878 →我们首先LF
89032 -4.05647 →我们无论如何→-0.175264LF
89033 -4.04555 →我们谨慎→-0.175264LF
89034 -4.01343 →我们正确→-0.175264LF
89035 -3.65107 →我们老板LF
89036 -3.90187 →我们当LF
89037 -4.2301 →我们顿LF
89038 -3.17385 →我们听说→-0.31132LF
89039 -2.71077 →我们很快→-0.754989LF
89040 -4.19345 →我们不久LF
89041 -4.05547 →我们恳求LF
89042 -4.20533 →我们以后LF
```

431166 -1.02001 → 是 的 LE
431167 -1.06692 → 是 身体 上 LE
431168 -0.567157 → 是 身体 健康 LE
431169 -1.27857 → 是 身体 关节 LE
431170 -0.77255 → 是 经过 适当 LE
431171 -1.11653 → 是 关于 一个 LE
431172 -1.12168 → 是 关于 教会 LE
431173 -0.388953 → 是 是否 有 LE
431174 -0.204048 → 是 一切 。 LE
431175 -0.578119 → 是 500 美元 LE
431176 -0.0641712 → 是 荒谬 的 LE
431177 -1.09409 → 是 谁 的 LE
431178 -0.863096 → 是 谁 。 LE
431179 -0.51727 → 是 谁 ? LE
431180 -1.5328 → 是 谁 把 LE

To make the faster loading, use the command to binary the zh.arpa file:

```
wangyiming@lobo:~$ wangyiming@lobo:~$ wangyiming@lobo:~$ /smt/mosesdecoder/bin/build_binary Aaron/Moses/lm/CN/Oxford-dictionary.zh.ar  
ps Aaron/Moses/lm/CN/Oxford-dictionary.zh.blm
```

```
Connected to 161.64.89.186          SSH2 - aes128-cbc - hmac-md5 | 95x29 | NU  
angyiming@lobo:~$ /smt/mosesdecoder/bin/build_binary Aaron/Moses/lm/CN/Oxford-dictionary.zh.ar  
a Aaron/Moses/lm/CN/Oxford-dictionary.zh.blm  
reading Aaron/Moses/lm/CN/Oxford-dictionary.zh.arpa  
----5---10---15---20---25---30---35---40---45---50---55---60---65---70---75---80---85---90---95  
-100  
*****  
***  
JCESS  
angviming@lobo:~$
```

Remote Name		Size	Type	Modified	Attributes
Oxford-dictionary.zh.arpa		15,364,649	ARPA 文...	12/12/2013 08:2...	-rw-rw-r--
Oxford-dictionary.zh.blm		13,745,321	BLM 文件	12/12/2013 08:3...	-rw-r--r--
Oxford-dictionary.zh.lm.gz		4,701,426	WinRAR...	12/12/2013 08:1...	-rw-rw-r--
Oxford-dictionary.zh.sb		7,747,108	SB 文件	12/12/2013 08:0...	-rw-rw-r--
Oxford-dictionary.zh.tok		6,667,108	TOK 文件	03/19/2013 01:3...	-rw-rw-r--

If use the following command, it will show the following result:

```
wangyiming@lobo:~$ 
wangyiming@lobo:~$ 
wangyiming@lobo:~$ echo "我来学校了 i come to university" | /smt/mosesdecoder/bin/query Aaron/
ses/lm/CN/Oxford-dictionary.zh.blm
Loading statistics:
Name:query      VmPeak:30424 kB VmRSS:14460 kB RSSMax:14460 kB user:0.004      sys:0    CPU:0.0
04      real:0
$: command not found
After queries:
Name:query      VmPeak:30428 kB VmRSS:14460 kB RSSMax:14460 kB user:0.004      sys:0    CPU:0.0
04      real:0
Total time including destruction:
Name:query      VmPeak:30428 kB VmRSS:1532 kB RSSMax:14460 kB user:0.004      sys:0    CPU:0.0
04      real:0
wangyiming@lobo:~$
```

Corpus for Training translation model:

Use the corpus 200,000 PT-CN bilingual sentences extracted from online websites of Macau SAR.

Prepare PT corpus:

Parse the PT.200000.tok sentence using the trained grammar on all the Floresta-Bosque corpus (train+deve+test):

Train the grammar:

新建文件夹				
名称	修改日期	类型	大小	
split	12/13/2013 11:2...	文件夹		
00.mrg	12/13/2013 11:2...	MRG 文件	3,275 KB	
01.mrg	12/13/2013 11:0...	MRG 文件	328 KB	
02.mrg	12/13/2013 11:0...	MRG 文件	327 KB	

This time, the 00.mrg contains “train+deve+test” sentences, 01.mrg contains deve sentences, 02.mrg contains testing sentences.

Run Configurations

Create, manage, and run configurations

Run a Java application

Name: GrammarTester

Main Arguments JRE Classpath Source Environment Common

Project: BerkeleyParserFloresta

Main class: edu.berkeley.nlp.PCFGIA.GrammarTrainer

Include system libraries when searching for a main class

Include inherited mains when searching for a main class

Stop in main

Run Configurations

Create, manage, and run configurations

Run a Java application

Name: GrammarTester

Main Arguments JRE Classpath Source Environment Common

Program arguments: -path corpus-ori-oneline -out GraFlorestaOri-allCorpus -treebank FLORESTA

VM arguments:

Working directory: \$workspace_loc:/BerkeleyParserFloresta

GrammarTester [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Dec 13, 2013, 11:36:27 AM)

```

Calling with { -path => corpus-ori-oneline -out => GraFlorestaOri-allCorpus -treebank => FLORESTA }
Loading trees from corpus-ori-oneline and using language FLORESTA
Will remove sentences with more than 10000 words.
Using horizontal=0 and vertical=1 markovization.
Using RIGHT binarization.
Using a randomness value of 1.0
Using grammar output file GraFlorestaOri-allCorpus.
Random number generator seeded at 2.
I will do at least 50 iterations.
Using smoothing parameters 0.5 and 0.1
Loading FLORESTA data!
Loading Floresta treebank trees...9289 957 957 939 trees...done
In training set we have # of words: 210839
reducing number of training trees from 9289 to 9289
Binarizing and annotating trees...

```

Finished as:

Problems @ Javadoc Declaration Console
 <terminated> GrammarTester [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (Dec 13, 2013, 11:36:27 AM)
 Calculating validation likelihood...done: -126357.39109627438
 Calculating training likelihood...done: -1267915.9627348997
 Beginning iteration 7:
 Calculating validation likelihood...done: -126354.85533110549
 Calculating training likelihood...done: -1267870.8670421112
 Beginning iteration 8:
 Calculating validation likelihood...done: -126351.21003701226
 Calculating training likelihood...done: -1267816.699571391
 Beginning iteration 9:
 Calculating validation likelihood...done: -126346.50682392207
 Calculating training likelihood...done: -1267757.4233067557
 Saving grammar to GraFlorestaOri-allCorpus_5_smoothing.gr.
 Saving successful.
 Calculating last validation likelihood...done.
 Iteration 10 (final) gives validation likelihood -126341.00721936491
 Saving grammar to GraFlorestaOri-allCorpus.
 It gives a validation data log likelihood of: -126341.00721936491
 Saving successful.

□	GraFlorestaOri-allCorpus	12/13/2013 11:59 AM	文件	2,645 KB
□	GraFlorestaOri-allCorpus_1_merging.gr	12/13/2013 11:37 AM	GR 文件	500 KB
□	GraFlorestaOri-allCorpus_1_smoothing.gr	12/13/2013 11:37 AM	GR 文件	499 KB
□	GraFlorestaOri-allCorpus_1_splitting.gr	12/13/2013 11:37 AM	GR 文件	529 KB
□	GraFlorestaOri-allCorpus_2_merging.gr	12/13/2013 11:38 AM	GR 文件	601 KB
□	GraFlorestaOri-allCorpus_2_smoothing.gr	12/13/2013 11:38 AM	GR 文件	618 KB
□	GraFlorestaOri-allCorpus_2_splitting.gr	12/13/2013 11:38 AM	GR 文件	703 KB
□	GraFlorestaOri-allCorpus_3_merging.gr	12/13/2013 11:40 AM	GR 文件	764 KB
□	GraFlorestaOri-allCorpus_3_smoothing.gr	12/13/2013 11:40 AM	GR 文件	859 KB
□	GraFlorestaOri-allCorpus_3_splitting.gr	12/13/2013 11:39 AM	GR 文件	978 KB
□	GraFlorestaOri-allCorpus_4_merging.gr	12/13/2013 11:45 AM	GR 文件	1,041 KB
□	GraFlorestaOri-allCorpus_4_smoothing.gr	12/13/2013 11:45 AM	GR 文件	1,374 KB
□	GraFlorestaOri-allCorpus_4_splitting.gr	12/13/2013 11:44 AM	GR 文件	1,511 KB
□	GraFlorestaOri-allCorpus_5_merging.gr	12/13/2013 11:58 AM	GR 文件	1,611 KB
□	GraFlorestaOri-allCorpus_5_smoothing.gr	12/13/2013 11:59 AM	GR 文件	2,651 KB
□	GraFlorestaOri-allCorpus_5_splitting.gr	12/13/2013 11:55 AM	GR 文件	2,560 KB

Put the corpus and the grammar in the 186-server for parsing:

Add |

 /home/wangyiming/Aaron/PT-CN
 Add

	Remote Name	Size	Type	Modified
	200000text.pt.tok	30,811,591	TOK 文件	12/13/2013 04:1
	BerkeleyParser-1.7.jar	3,092,739	Executa...	10/17/2013 07:3
	GraFlorestaOri-allCorpus	2,707,594	文件	12/13/2013 11:5

The parsing command, the parsing on server-186 began at 20131213-4:17 PM, finished at 20131215-03:48:36AM

```
wangyiming@lobo:~/Aaron/PT-CN$  
wangyiming@lobo:~/Aaron/PT-CN$ java -jar BerkeleyParser-1.7.jar -gr GraFlorestaO  
ri-allCorpus -inputFile 200000text.pt.tok -outputFile 200000text.pt.tok.parsed.B  
osque
```

Connected to 161.64.89.186

SSH2 - aes128-cbc - hmac-md5 | 80x24

Parsing finished as:

1:161.64.89.186 - Aaron20131213 - SSH Secure Shell

File Edit View Window Help

Quick Connect Profiles

SSH Secure Shell 3.2.0 (Build 267)
Copyright (c) 2000-2002 SSH Communications Security Corp - <http://www.ssh.com/>

This copy of SSH Secure Shell is licensed for educational, charity,
or personal recreational or hobby use.
Any commercial use requires a separate license.

Welcome to Ubuntu 12.04.2 LTS (GNU/Linux 3.5.0-30-generic x86_64)

* Documentation: <https://help.ubuntu.com/>

System information disabled due to load higher than 24.0

*** System restart required ***

Last login: Thu Dec 12 16:29:29 2013 from 161.64.89.126
wangyiming@lobo:~\$ cd Aaron/
wangyiming@lobo:~/Aaron\$ cd PT-CN/
wangyiming@lobo:~/Aaron/PT-CN\$
wangyiming@lobo:~/Aaron/PT-CN\$ java -jar BerkeleyParser-1.7.jar -gr GraFlorestaO
ri-allCorpus -inputFile 200000text.pt.tok -outputFile 200000text.pt.tok.parsed.B
osque

wangyiming@lobo:~/Aaron/PT-CN\$

<http://www.ssh.com/> SSH2 - aes128-cbc - hmac-md5 | 80x28

Remote Name		Size	Type	Modified	Attributes
200000text.pt.tok		30,811,591	TOK 文件	12/13/2013 04:12:22 PM	-rw-rw-r--
200000text.pt.parsed.Bosque		84,690,309	BOSQU...	12/15/2013 03:48:36 AM	-rw-rw-r--
BerkeleyParser-1.7.jar		3,092,739	Execut...	10/17/2013 07:38:58 PM	-rw-rw-r--
GraFlorestaOri-allCorpus		2,707,594	文件	12/13/2013 11:59:18 AM	-rw-rw-r--

For the uni-nofuntag-Floresta-Bosque setting: We should replace the ori-phrase tags parsed from last step with the uni-phrase. (If we use the uni-phrase-grammar-trained to parse the plain.tok PT text to gain the uni-phrase taggest corpus, it is not suitable because the grammar-trained-pt-floresta and the grammar-trained-pt-floresta-uniphrase have different accuracies.)

Prepare the CN corpus:

We parse the 200,000 chinese.tok sentences using the GrammarTrainedCTB-7; parsing began at 20131213-10:47AM, finished at 20131214-04:18 AM;

```
E:\Berkeley_Parser>
E:\Berkeley_Parser>
E:\Berkeley_Parser>java -jar BerkeleyParser-1.7.jar -gr GrammarTrained-CTB7 -inputFile 200000text.zh.tok -outputFile 200000text.zh.tok.parsed
```

The training finished as:

```
E:\Berkeley_Parser>
E:\Berkeley_Parser>
E:\Berkeley_Parser>java -jar BerkeleyParser-1.7.jar -gr GrammarTrained-CTB7 -inputFile 200000text.zh.tok -outputFile 200000text.zh.tok.parsed

E:\Berkeley_Parser>
```

200000text.zh.tok	12/12/2013 4:17 PM	TOK 文件	25,950 KB
200000text.zh.tok.parsed	12/14/2013 4:18 AM	PARSED 文件	74,607 KB
BerkeleyParser-1.7.jar	10/17/2013 7:39 PM	Executable Jar File	3,021 KB
chn_sm5.gr	10/17/2013 7:39 PM	GR 文件	15,965 KB
chtb_0001.nw.seg	9/18/2010 10:48 AM	SEG 文件	3 KB
commands from Anson.txt	10/18/2013 10:47 AM	文本文档	1 KB
eng_sm6.gr	10/17/2013 7:38 PM	GR 文件	21,722 KB
englishout.txt	10/18/2013 10:56 AM	文本文档	1 KB
GrammarTrained-CTB7	11/12/2013 3:52 PM	文件	11,331 KB

NMT05fr-en-dev2000.fr.tok.pt | europarl-v7/en-fr.tok.pt | NMT05fr-en-test2000.fr.tok.pt | europarl-v7/en-fr.tok.parsed.pt | 200000ext.zh.tok.parsed.pt |

```

1 ((IP (FU '')) (IP (NP (QP (CD 一)) (NP (NN 朝)))) (VP (LB 被)) (IP (NP (NN 蛇)) (VP (VV 吸)))))) (FU ,) ((IP (VP (NP (NT 十年)) (VP (VV 怕)) (IP (VP (VV 草绳)))))) (FU '')) )■
2 ((IP (VP (VP (ADVP (AD 又)) (VP (VV 妻)) (VP (ADVP (AD 马儿)) (VP (VV 道)))))) (FU ,) (VP (ADVP (AD 又)) (VP (VV 妻)) (VP (VP (VV 马儿)) (VP (ADVP (AD 不)) (VP (VV 吃)) (NP (NN 草
3 ((IP (VP (VV 进行)) (NP (DNP (NP (NP (NN 纪律)) (FU ,) (NP (ADJP (JJ 全面)) (NP (NN 调查)))) (CC 及) (NP (DNP (NP (ADJP (JJ 简易)) (NP (NN 调查)))) (CC
4 ((IP (NP (NN 总督)) (NP (NN 办公室)) (VP (VC 即)) (NP (NN 政府)) (NN 司) (NN 办公室)) (NN 行政) (NN 辅助) (NN 部门)))) )■
5 ((IP (VP (VV 运用)) (NP (CP (IP (NP (ADZJ (JJ 公共)) (NP (NN 部队)))) (VP (VV 勉善)) (NP (NP (NN 法律))) (CC 或) (NP (ADJP (JJ 正当)) (NP (NN 命令)))))) (DEC 的)) (NP (NN 执行)))) 
6 ((IP (VP (VP (ADVP (AD 又)) (VP (VV 妻)) (VP (ADVP (AD 马儿)) (VP (VV 道)))))) (CC 和) (NP (DP (DT 其他)) (CP (IP (VP (VA 残忍)) (FU ,) (VA 不人道)) (CC 或) (VP (VE 有)) (NP (NN 妻)) (NN 人格)))))) (DEC 的)) (NP
7 ((NP (CP (IP (VP (VV 关)) (PP (P 为)) (IP (VP (VV 制止)) (IP (VP (VV 危害)) (NP (DNP (NP (ADJP (JJ 民用)) (NP (NN 航空)) (NN 安全)))) (DEG 的)) (ADJP (JJ 非法)) (NP (NN 行为))))))) 
8 ((IP (VP (VV 禁止)) (IP (VP (VP (VV 贩卖)) (NP (NN 人口)))) (CC 及) (VP (VP (VV 取缔)) (NP (NN 营利))) (VP (VV 使)) (NP (NN 人)) (IP (VP (VV 姦淫)) (NP (NN 公约))))))) )■
9 ((IP (VP (VP (VV 关)) (PP (P 为)) (NP (DNP (PP (P 在)) (LCP (NP (NN 航空器)) (LC 内)))) (DEG 的)) (NP (NN 犯罪)))) (CC 和) (VP (VV 犯)) (NP (CP (IP (VP (VE 有)) (NP (DP (DT 某些)) (O
0 ((NP (CP (IP (VP (VV 关)) (PP (P 从)) (NP (NN 国外)))) (VP (VV 调取)) (NP (NN 民事)) (CC 或) (NN 商事) (NN 证据))))))) (DEC 的)) (NP (NN 公约))) )■
1 ((IP (VP (VV 须)) (VP (PP (P 透过)) (IP (NP (NN 计算)) (FU ,) (NN 盈))) (VP (VA 重)) (CC 或) (VA 量度)))) (VP (VV 予以)) (NP (DNP (ADJP (JJ 确定)) (DEG 的)) (NP (NN 动产))))))) )■
2 ((NP (DNP (LCP (IP (VP (LB 被)) (IP (NP (NN 继承人)) (VP (VP (VV 作出)) (NP (NN 意思)))) (VP (VV 表示))))))) (LC 时)) (DEG 的)) (IP (VP (VV 属)) (NP (NN 人)))) (NP (NN 法))) )■

```

After the parsing finish of the CN corpus, we replace the ori-phrase-tags using the uni-nofuntag-phrase-tags

Training translation model use ori-tags:

Convert the 200,000 PT and CN parsed sentences into mosesxml format.

Test translation model sue ori-tags:

Training translation model use uni-tags:

Convert the 200,000 PT and CN parsed sentences into mosesxml format.

Test translation model sue uni-tags:

Tune the MT system using LEPOR metric (external metric):

1. Change the xxxx file as below to call the LEPOR metric.

2. Change the LEPOR source document as below:

3. Put the source document of LEPOR metric at yyy

4. Run the system like below to test

5.

=====

Reference:

Moses manual: [<http://www.statmt.org/moses/manual/manual.pdf>], accessed 2013.10.31

TianLiang' blog: [http://www.tianliang123.com/moses_installation], accessed 2013.10.31.