```
######
# Fichier                     : 3 étapes de faire marcher MOSES
# Auteur                      : Tan LE
# Date de création            : samedi, 19/03/2016
# Date de modification        : samedi, 12/10/2019

######
```

VirtualBox VM version 4.3.24
    Ubuntu 14.04 LTS (64 bits)
        Ram : 9 Go
        i5, 2.67 GHz
        266.60 Go Disque dur

Question :: How to install Moses on Ubuntu 64 bits ?
Sources ::
    http://www.achchuthan.org/2014/06/install-moses-on-ubuntu-14.04.html
    http://www.statmt.org/moses/?n=Moses.Baseline
    http://www.statmt.org/moses/?n=Development.GetStarted

STEP 1 :: PREPROCESSING
    Muc dich :
        Tokenize, Lowercase, truecase, Clean cutoff 1-100

```
        ~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en
news-commentary-v11.fr-en.en > news-commentary-v11.fr-en.tok.en

        ~/mosesdecoder/scripts/tokenizer/tokenizer.perl -l fr
news-commentary-v11.fr-en.fr > news-commentary-v11.fr-en.tok.fr

        ~/mosesdecoder/scripts/training/clean-corpus-n.perl corpus/FrEn/test.tok.true fr
en corpus/FrEn/test.clean 1 80

         ~/mosesdecoder/scripts/training/clean-corpus-n.perl
~/SMT-FR-EN/corpus_for_training/hansard.house.debate_preprocessed_script_2 fr en
~/SMT-FR-EN/corpus_for_training/hansard.house.debate_preprocessed_script_2.clean 1 100
clean-corpus.perl: processing
/home/tan/SMT-FR-EN/corpus_for_training/hansard.house.debate_preprocessed_script_2.fr &
.en to
/home/tan/SMT-FR-EN/corpus_for_training/hansard.house.debate_preprocessed_script_2.clea
n, cutoff 1-100, ratio 9
..........(100000)..........(200000)..........(300000)..........(400000)..........(500000)..........(600000)......
```

```
....(700000)......
Input sentences: 761017  Output sentences:  760162
```

Tao Language Model (SRILM) :: langue cible EN for example FR-EN SMT

```
## SRILM
cd srilm/bin/i686-m64/
./ngram-count -order 4 -text ~/corpus/FrEn/training.clean.en -lm
~/lm/training.fr-en.srilm.arpa.en

~/mosesdecoder/bin/build_binary training.fr-en.srilm.arpa.en
training.fr-en.srilm.blm.en
```

STEP 2 :: TRAIN MODEL

Muc dich :

chay MGIZA

tao Translation Model

```
## cau lenh train-model : su dung LM binary, MGIZA
nohup nice /home/tan/mosesdecoder/scripts/training/train-model.perl -root-dir
/home/tan/SMT_Projects/works/workFREN_wmt13 -corpus
/home/tan/SMT_Projects/works/workFREN_wmt13/corpus/training.clean -f fr -e en -mgiza
-mgiza-cpus 4 -parallel -core 8 -reordering msd-bidirectional-fe -lm
0:3:/home/tan/SMT_Projects/works/workFREN_wmt13/lm/tokenization.europarl-v7.fr-en.srilm.bl
m.en -external-bin-dir /home/tan/mosesdecoder/tools >&
/home/tan/SMT_Projects/works/workFREN_wmt13/training.out
```

STEP 3 :: RUN DECODER

Muc dich : tinh cac metrics NIST va BLEU theo cac tap tin *.SGM

```
## chay decoder de dich ra EN
nohup nice ~/mosesdecoder/bin/moses -f
~/SMT_Projects/works/workFREN_wmt13/model/moses.ini <
~/SMT_Projects/works/workFREN_wmt13/corpus/test_requeteRIT_fr_projet2_DIC9320.baselin
e.fr > ~/SMT_Projects/works/workFREN_wmt13/evaluation/test2.translated.output 2>
~/SMT_Projects/works/workFREN_wmt13/evaluation/decode.out

## mot cach tinh diem BLEU nhanh
~/mosesdecoder/scripts/generic/multi-bleu.perl -lc
~/SMT_Projects/works/workFREN_wmt13/corpus/test_requeteRIT_fr_projet2_DIC9320.baselin
e.en < ~/SMT_Projects/works/workFREN_wmt13/evaluation/test2.translated.output >
```

```
SMT_Projects/works/workFREN_wmt13/evaluation/bleu.out


# tao cac tap tin *.SGM
## output :: EN
perl /home/tan/SMT_Projects/_tools/scripts/my_wrapxml.perl tstset fr en
/home/tan/SMT_Projects/works/workFREN_wmt13/evaluation/corpusFREN_wmt13.output
/home/tan/SMT_Projects/works/workFREN_wmt13/evaluation/corpusFREN_wmt13.output.sgm

## source :: FR
perl /home/tan/SMT_Projects/_tools/scripts/my_wrapxml.perl srcset fr en
/home/tan/SMT_Projects/works/workFREN_wmt13/corpus/test_requeteRIT_fr_projet2_DIC9320
.fr
/home/tan/SMT_Projects/works/workFREN_wmt13/evaluation/corpusFREN_wmt13-src.fr.sgm


## ref :: EN
perl /home/tan/SMT_Projects/_tools/scripts/my_wrapxml.perl refset fr en
/home/tan/SMT_Projects/works/workFREN_wmt13/corpus/test_requeteRIT_fr_projet2_DIC9320
.en
/home/tan/SMT_Projects/works/workFREN_wmt13/evaluation/corpusFREN_wmt13-ref.en.sgm

/home/tan/SMT_Projects/_tools/mteval-v13a-20091001/mteval-v13a.pl -s
/home/tan/SMT_Projects/works/workFREN_wmt13/evaluation/corpusFREN_wmt13-src.fr.sgm -r
/home/tan/SMT_Projects/works/workFREN_wmt13/evaluation/corpusFREN_wmt13-ref.en.sgm
-t
/home/tan/SMT_Projects/works/workFREN_wmt13/evaluation/corpusFREN_wmt13.output.sgm
-c > /home/tan/SMT_Projects/works/workFREN_wmt13/report.out


## analyser les OOV
# check in language model
cat works/evaluation/test.preprocessed.en | ~/mosesdecoder/bin/query
LM/tokenization.europarl-v7.fr-en.srilm.blm.en > OOV_analysis_query_LM_1M_pairs_courtes

# check in translation model :: training data SOURCE & CIBLE
cat ~/tan_SMT_Projet2_DIC9320/train.preprocessed.fr | ~/mosesdecoder/scripts/analysis/oov.pl
~/tan_SMT_Projet2_DIC9320/test.preprocessed.fr >
~/tan_SMT_Projet2_DIC9320/OOV_analysis_FR_avec_script_OOV.pl.out

cat ~/tan_SMT_Projet2_DIC9320/train.preprocessed.en |
~/mosesdecoder/scripts/analysis/oov.pl ~/tan_SMT_Projet2_DIC9320/test.preprocessed.en >
~/tan_SMT_Projet2_DIC9320/OOV_analysis_EN_avec_script_OOV.pl.out
```

```
# script de dich
cd tan_SMT_news.v11
tan@tan-VirtualBox:~/tan_SMT_news.v11$ perl ~/translate.pl test.preprocessed.fr
model/moses.ini . --local

# script de tinh diem bleu
perl ~/multi-bleu.perl '/home/tan/tan_SMT_news.v11/test.preprocessed.en' <
'/home/tan/tan_SMT_news.v11/moses.output'
BLEU = 9.02, 46.6/22.0/8.9/0.7 (BP=1.000, ratio=1.265, hyp_len=277, ref_len=219)
```

Experimentations ::

FR-EN :

20k
training set = 80% = 18k
dev set = 10% = 1k
test set = 10% = 1k

cleaning : cutoff 1-80 ratio=9

training set = 17.803 sentences pairs
test set = 992 sentences pairs

Time of process : 15 minutes

```
BLEU = 19.25, 54.1/25.0/13.3/7.6 (BP=1.000, ratio=1.113, hyp_len=27863,
ref_len=25035)
```