Machine Translation
Homework 03 - Preprocessing

Aqeel Labash

November 16, 2016

1 Introduction

For this home work I used `Opensubtitle` to translate from `Arabic` to `English`. The original file contained $16000,000 \sim$ sentence. I used `Trainin: 1000000`, `dev:5000` and `test:2000`.

2 General Code

```
MOSES=/home/aqeel/MT/moses
echo Preparing Data
tst=2000
trn=1000000
dev=5000
all=0
let "all=tst+trn+dev"

echo Take $trn for training , $dev for dev , $tst for testing
sed -n 1,"$all"p OpenSubtitles2016.ar-en.ar > all.ar
sed -n 1,"$all"p OpenSubtitles2016.ar-en.en > all.en
echo Split the data
lstart=0
lend=0
let "stlimit+=1"
let "lend=trn"
sed -n "$stlimit","$lend"p all.ar > train.ar
sed -n "$stlimit","$lend"p all.en > train.en
let "stlimit+=trn"
let "lend=stlimit+dev"
sed -n "$stlimit","$lend"p all.ar > dev.ar
sed -n "$stlimit","$lend"p all.en > dev.en

let "stlimit+=dev"
let "lend = stlimit+tst"
sed -n "$stlimit","$lend"p  all.ar > test.ar
sed -n "$stlimit","$lend"p all.en > test.en
rm all.ar all.en
echo Tokinizing Data
for f in {train,dev,test}.{ar,en};
do
  $MOSES/scripts/tokenizer/tokenizer.perl -threads 8 < $f > tok-$f
  rm $f
done
echo Truecasing English
$MOSES/scripts/recaser/train-truecaser.perl --model en-truecase.mdl --corpus tok-train.en
for f in {test,dev,train}.en;
do
  $MOSES/scripts/recaser/truecase.perl --model en-truecase.mdl < tok-$f > tc-tok-$f
  rm tok-$f
done
for f in {train,dev,test};
do
  mv tok-$f.ar tc-tok-$f.ar
  rm tok-$f.ar
done

$MOSES/+scripts/training/clean-corpus-n.perl tokenized-and-lowecased ar en cleaned 1 100

$MOSES/bin/lmplz -o 5 -S 50% < tc-tok-train.en > lm-en.arpa
```

```
$MOSES/bin/build_binary lm-en.arpa lm-en.blm

$MOSES/scripts/training/train-model.perl --corpus tc-tok-train --f ar --e en --external-bin-dir $MOSES/bin --lm 0:5:$(pwd)/lm-en.blm --root-dir mt-experiment-1 --reordering msd-bidirectional-fe --mgiza --mgiza-cpus 8
# To Enhance Peroframce Later (on Testing level).
cd mt-experiment-1
mkdir binarised-model
$MOSES/bin/processPhraseTableMin -in model/phrase-table.gz -nscores 4 -out binarised-model/phrase-table
$MOSES/bin/processLexicalTableMin -in model/reordering-table.wbe-msd-bidirectional-fe.gz -out binarised-model/reordering-table
cd ..


for i in `seq 1 3`;
do
  echo ======================$i======================
# $MOSES/scripts/training/mert-moses.pl $(pwd)/tc-tok-dev.ar $(pwd)/tc-tok-dev.en $MOSES/bin/moses train/model/moses.ini --mertdir $MOSES/bin/ --decoder-flags="-threads all" &> mert$i.out
  $MOSES/scripts/training/mert-moses.pl $(pwd)/tc-tok-dev.ar $(pwd)/tc-tok-dev.en $MOSES/bin/moses $(pwd)/mt-experiment-1/model/moses.ini --working-dir $(pwd)/mt-experiment-1/mert-$i --threads 4 --decoder-flags "--threads 4" > mert-$i.out
done
```

Since I used `cmph` to decrease the amount of memory required later, I had to split the bash file to update the `Path` value in `moses.ini` in `mert` files.[1]
The second part is to calculate the `BLEU` score.

```
for i in `seq 1 3`;
do
  $MOSES/bin/moses -f mt-experiment-1/mert-$i/moses.ini -i tc-tok-test.ar > mt-experiment-1/mert-$i/hypothesis0.ar
  $MOSES/scripts/generic/multi-bleu.perl tc-tok-test.en < mt-experiment-1/mert-$i/hypothesis0.ar > out_$i.txt
done
```

3 The Random Sentences

For random sentences I just ran a the following code :

```
import random
for i in range(10):
  print random.randint(0,2001)
```

Then I enhanced the chosen sentences (to decrease the number of small sentences). in Figure 1 we can see the sentences[2].

[1] Thanks To Maksym, didn't know about it before.
[2] I used images to put arabic text due to compatability issue between arabic and latex and my needs.

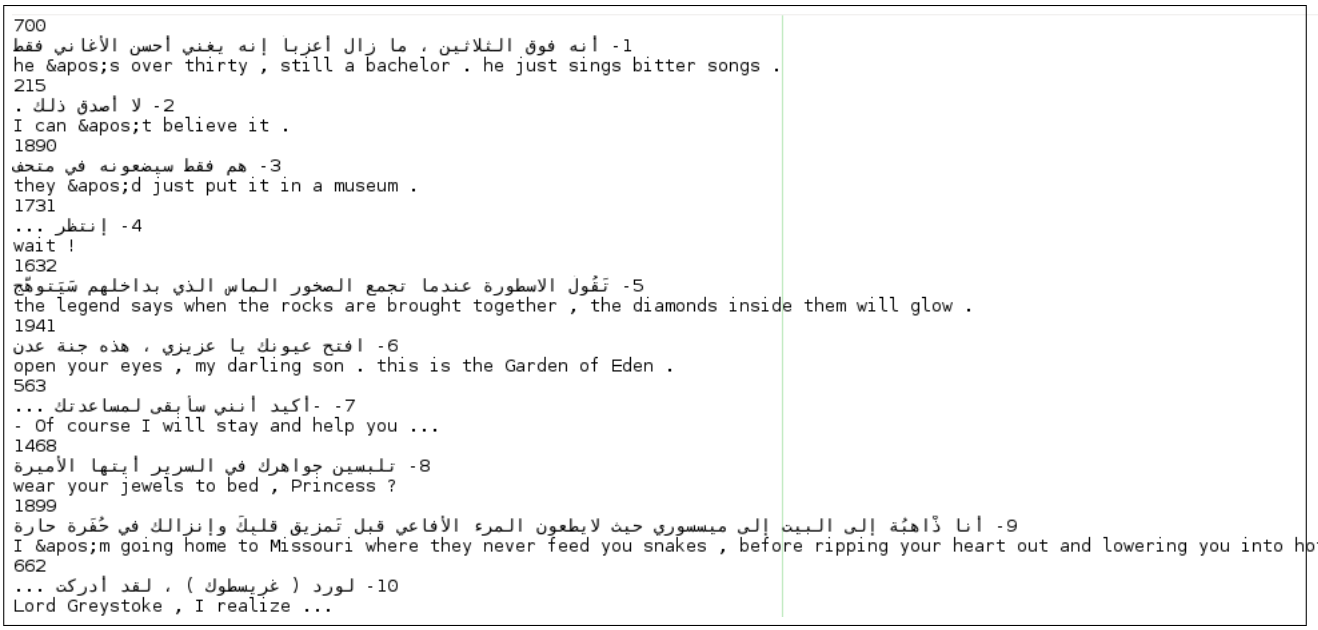

Fig. 1: the sentence number in random file, followed by Arabic sentence (right to left) starting with sentence number (1 to 10) followed by the english sentence.(left to right)

4 Notes on the original translation

- I would say the translation it self is not accurate in sense of using different expresion (using best instead of better) and in sense of the using mix of accents with standard accent.
- The translation it self is not the best form that I would use to translate those seneteces.
- diarectecs is also included in some cases (removing them lead to lose of meaning in some cases).

5 Baseline

For the baseline (used the previous bash code exactly) I got the following score[3] score:
`BLEU = 24.23, 57.5/31.8/19.5/12.1 (BP=0.945, ratio=0.946, hyp_len=13502, ref_len=14266)`
Figure 2 show the baseline translation.

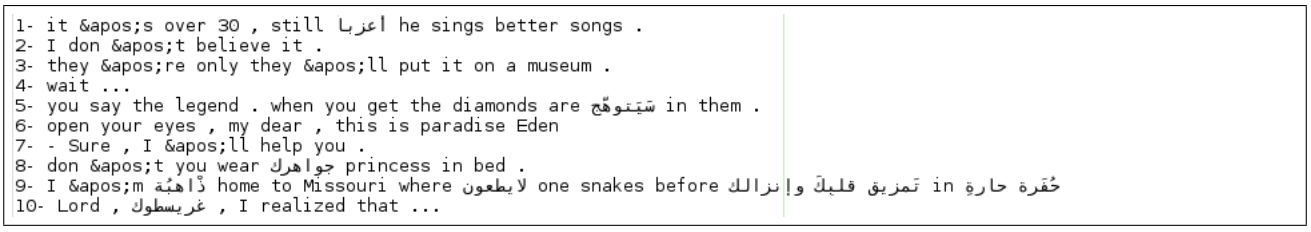

Fig. 2: baseline translation with id

6 Compund Split

For this method I used the bash code here. The best score was :
`BLEU = 23.81, 57.6/31.8/19.3/11.9 (BP=0.934, ratio=0.936, hyp_len=13359, ref_len=14266)`
The output for the same sentences where as shown in Figure 3 3

[3]The best score between all the 3 tuning operations

```
Compound Spliter
1- it &apos;s over 30 , still أعزبا he sings better songs .
2- I don &apos;t believe it .
3- they &apos;re only they &apos;ll put it on a museum .
4- wait ...
5- you say legend . when the rocks سَيَتوهّج diamonds are in them .
6- open your eyes , my dear , this is the Garden of
7- - Sure , I &apos;ll help you .
8- don &apos;t you wear جواهرك princess in bed .
9- I &apos;m ذّاهبُة home to Missouri where لايطعون one snakes before تَمزيق قلبكَ وإنزالك in حُفَرة حارةٍ
10- Lord , غريسطوك , I realized that ...
```

Fig. 3: compound splitter translation with id

7 Byte Pair Encoding

After 26 hours of computation time I found that it has a bug with arabic language (instead of output like `a@@` → `a @ @`)

8 Comparision

First let's take a look at Figure 4 were we can see all the data in one place.

```
===== ALL INFO====
(1)
AR- أنه فوق الثلاثين ، ما زال أعزبا إنه يغني أحسن الأغاني فقط
OR- he &apos;s over thirty , still a bachelor . he just sings bitter songs .
BL- it &apos;s over 30 , still أعزبا he sings better songs .
CS- it &apos;s over 30 , still أعزبا he sings better songs .
(2)
AR- لا أصدق ذلك .
OR- I can &apos;t believe it .
BL- I don &apos;t believe it .
CS- I don &apos;t believe it .
(3)
AR- هم فقط سيضعونه في متحف
OR- they &apos;d just put it in a museum .
BL- they &apos;re only they &apos;ll put it on a museum .
CS- they &apos;re only they &apos;ll put it on a museum .
(4)
AR- إنتظر ...
OR- wait !
BL- wait ...
CS- wait ...
(5)
AR- تَقُول الاسطورة عندما تجمع الصخور الماس الذي بداخلهم سَيَتوهّج
OR- the legend says when the rocks are brought together , the diamonds inside them will glow .
BL- you say the legend . when you get the diamonds are سَيَتوهّج in them .
CS- you say legend . when the rocks سَيَتوهّج diamonds are in them .
(6)
AR- افتح عيونك يا عزيزي ، هذه جنة عدن
OR- open your eyes , my darling son . this is the Garden of Eden .
BL- open your eyes , my dear , this is paradise Eden
CS- open your eyes , my dear , this is the Garden of
(7)
AR- -اكيد انني سأبقى لمساعدتك ...
OR- - Of course I will stay and help you ...
BL- - Sure , I &apos;ll help you .
CS- - Sure , I &apos;ll help you .
(8)
AR- تلبسين جواهرك في السرير أيتها الأميرة
OR- wear your jewels to bed , Princess ?
BL- don &apos;t you wear جواهرك princess in bed .
CS- don &apos;t you wear جواهرك princess in bed .
(9)
AR- أنا ذّاهبُة إلى البيت إلى ميسسوري حيث لايطعون المرء الأفاعي قبل تَمزيق قلبكَ وإنزالك في حُفَرة حارةٍ
OR- I &apos;m going home to Missouri where they never feed you snakes , before ripping your heart out and lowering you into hot pits !
BL- I &apos;m ذّاهبُة home to Missouri where لايطعون one snakes before تَمزيق قلبكَ وإنزالك in حُفَرة حارةٍ
CS- I &apos;m ذّاهبُة home to Missouri where لايطعون one snakes before تَمزيق قلبكَ وإنزالك in حُفَرة حارةٍ
(10)
AR- لورد ( غريسطوك ) ، لقد أدركت ...
OR- Lord Greystoke , I realize ...
BL- Lord , غريسطوك , I realized that ...
CS- Lord , غريسطوك , I realized that ...
```

Fig. 4: show different translation, `AR: arabic, ORG: original translation, BL: Baseline, CS: Compound splitter`

The following table show the sentences and the methods:

Sentence ID	Baseline	Compound Split	Notes
1	same	same	Considered the error bitter and (it's instead of he) to equalize them
2	better	better	
3	worse	worse	
4	same	same	
5	worse	worse	
6	Better	Worse	
7	worse	worse	Even the original one is not that good :(
8	worse	worse	
9	worse	worse	(there is a typo in the original translation)
10	same	same	equalized the untranslated word with catching the right sentence time (past)

9 Comments

I would like to comment the following points :

- Using compound split didn't improve the performance. The result wasn't shocking since arabic language don't have much of long words (can't think of any long words like german words) unless we include the diaractects (which I suspect to be the difference between the baseline and compound splitter way).
- Looking at the random results I found problems in English subtitles and Arabic subtitles as well.
- The data mainly was English to Arabic but I used it in reverse. So I considered the Arabic text as the original and I liked that at some points the models performed better than the original English text in catching the sentence time.
- In the data there were some typos which popup the problem of unseen words.

10 Acknowledgement

Thanks to Hasan helped me to recognize the error when merging `migiza` with `moses`.
Thanks to Maksym, told me about `cmph` library which without it my laptop won't handle all this data.
Please Note: All tex,pdf,.sh,... files exist on github

E.O.F