

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 ~ sentence. I used `Trainin: 1000000, dev:5000 and test:2000`.

2 General Code

```
1 MOSES=/home/ aqeel /MT/ moses
2 echo Preparing Data
3 tst=2000
4 trn=1000000
5 dev=5000
6 all=0
7 let "all=tst+trn+dev"
8
9 echo Take $trn for training , $dev for dev , $tst for testing
10 sed -n 1,"$all"p Opensubtitles2016.ar-en.ar > all.ar
11 sed -n 1,"$all"p Opensubtitles2016.ar-en.en > all.en
12 echo Split the data
13 lstart=0
14 lend=0
15 let "stlimit+=1"
16 let "lend=trn"
17 sed -n "$stlimit","$lend"p all.ar > train.ar
18 sed -n "$stlimit","$lend"p all.en > train.en
19 let "stlimit+=trn"
20 let "lend=stlimit+dev"
21 sed -n "$stlimit","$lend"p all.ar > dev.ar
22 sed -n "$stlimit","$lend"p all.en > dev.en
23
24 let "stlimit+=dev"
25 let "lend = stlimit+tst"
26 sed -n "$stlimit","$lend"p all.ar > test.ar
27 sed -n "$stlimit","$lend"p all.en > test.en
28 rm all.ar all.en
29 echo Tokenizing Data
30 for f in {train,dev,test}.$ar,$en;
31 do
32     $MOSES/scripts/tokenizer/tokenizer.perl -threads 8 < $f > tok-$f
33     rm $f
34 done
35 echo Truecasing English
36 $MOSES/scripts/recaser/train-truecaser.perl --model en-truecase.mdl --corpus tok-train.en
37 for f in {test,dev,train}.en;
38 do
39     $MOSES/scripts/recaser/truecase.perl --model en-truecase.mdl < tok-$f > tc-tok-$f
40     rm tok-$f
41 done
42 for f in {train,dev,test};
43 do
44     mv tok-$f.$ar tc-tok-$f.$ar
45     rm tok-$f.$ar
46 done
47
48 $MOSES/scripts/training/clean-corpus-n.perl tokenized-and-lowecased ar en cleaned 1 100
49
50 $MOSES/bin/lmplz -o 5 -S 50% < tc-tok-train.en > lm-en.arpa
```

```

51 $MOSES/bin/build_binary lm-en.arpa lm-en.blm
52
53 $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
54 # To Enhance Peroframce Later (on Testing level).
55 cd mt-experiment-1
56 mkdir binarised-model
57 $MOSES/bin/processPhraseTableMin -in model/phrase-table.gz -nscores 4 -out binarised-model/
    phrase-table
58 $MOSES/bin/processLexicalTableMin -in model/reordering-table.wbe-msd-bidirectional-fe.gz -out
    binarised-model/reordering-table
59 cd ..
60
61
62 for i in `seq 1 3`;
63 do
64     echo =====${i}=====
65 # $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
66 $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
67 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.

```

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

```

3 The Random Sentences

For random sentences I just ran a the following code :

```

1 import random
2 for i in range(10):
3     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 sentences².

¹Thanks To Maksym, didn't know about it before.

²I used images to put arabic text due to compatability issue between arabic and latex and my needs.

700 1- ابنه فوق الثلاثين ، ما زال أعربا إنه يغني أحسن الأغاني فقط he 's over thirty , still a bachelor . he just sings bitter songs . 215 2- لا أصدق ذلك . I can 't believe it . 1890 3- هم فقط سيضعونه في متحف they 'd just put it in a museum . 1731 4- إنتظر ... wait ! 1632 5- تقول الاسطورة عندما تجمع الصخور الماس الذي بداخلهم سَتَنوَّج the legend says when the rocks are brought together , the diamonds inside them will glow . 1941 6- افتح عيونك يا عزيزي ، هذه جنة عدن open your eyes , my darling son . this is the Garden of Eden . 563 7- اكيد انني سأبقى لمساعدتك ... - Of course I will stay and help you ... 1468 8- تلبسين جواهرك في السرير أيتها الأميرة wear your jewels to bed , Princess ? 1899 9- أنا ذَاهِيَّة إلى البيت إلى ميسسوري حيث لا يطعمون المرء الأفاعي قبل تمزيق قلبك وإنزالك في حُفرة حارة I 'm going home to Missouri where they never feed you snakes , before ripping your heart out and lowering you into hot pits ! 662 10- لورد (غريستوك) ، لقد أدركت ... Lord Greystoke , I realize ...	
---	--

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.
- diarctecs 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 score³ 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.

1- it 's over 30 , still أعربا he sings better songs . 2- I don 't believe it . 3- they 're only they 'll put it on a museum . 4- wait ... 5- you say the legend . when you get the diamonds are سَتَنوَّج in them . 6- open your eyes , my dear , this is paradise Eden 7- - Sure , I 'll help you . 8- don 't you wear جواهرك princess in bed . 9- I 'm ذَاهِيَّة home to Missouri where لا يطعمون one snakes before إنزالك 10- Lord , غريستوك , I realized that ...	
--	--

Fig. 2: baseline translation with id

6 Compound Split

7 Byte Pair Encoding

I used this way (took around one day per one tuning process and couldn't afford more time to invest

³The best score between all the 3 tuning operations