INF5820, Assignment 3: Machine Translation

Second part, fall 2012

2 Alignment

We will experiment with the well known alignment tool GIZA++. You will find the relevant material in /projects/nlp/external-bin-dir (which you may download to your own file area).

Part a

To familiarize ourselves with the tool, we will consider the simple example from the lecture. Make two files eng and nor with two lines each:

```
dog bit dog
dog barket
```

and

```
hund bet hund
hund bjeffeet
```

To prepare the corpus for GIZA++ we have to compute some additional files. Run first

```
./plain2snt.out nor eng
```

Take a look on the newly generated files and try to understand what they contain. An .vcb-file assigns a unique numerical identifier to each word. The second number, e.g., 3 in

```
2 hund 3
```

counts the number of ocurrences of this word. In the .snt files, the words in the source files are replaced by identifiers.

We also need one more initial preparation

```
./snt2cooc.out eng.vcb nor.vcb eng_nor.snt > corp.cooc
```

We are not concerned with the content of this file but it is used in the next step.

Part b

We can then run the alignment program

```
./GIZA++ -S eng.vcb -T nor.vcb -C eng_nor.snt \
-CoocurrenceFile corp.cooc
```

This generates a series of files with a longish prefix. You may use a prefix of your choice by the option "-o", e.g.,

```
./GIZA++ -S eng.vcb -T nor.vcb -C eng_nor.snt \
-CoocurrenceFile corp.cooc -o experiment1
```

To see what is going on and what the files contain we may take a look at the screen output from the run (which we could dump to a file by)

```
./GIZA++ -S eng.vcb -T nor.vcb -C eng_nor.snt \
-CoocurrenceFile corp.cooc -o experiment1 > output1
```

We see that the program by default has run 5 iterations of IBM model1 followed by 5 iterations of model3 and 5 itertions of model4.

The names of the produced files indicates

- whether it is an alignment file (a/A), a lexical probabilities file (t) etc.
- which IBM model has produced it
- and after how many iterations

For example "experiment1.t3.final" is the word probabilities after the program has finished model 3. To make this readable, we have made a script which works for small files.

```
./nums2words.py eng.vcb nor.vcb experiment.t3.final
```

Take a look at the produced file. Also have a look at "experiment1.a3.final" and try to understand it.

Part c

We may change the default values (which we see in output1). For example, to study the convergence behavior of Model1 we may iterate it a hundred times by

```
./GIZA++ -S eng.vcb -T nor.vcb -C eng_nor.snt \
-CoocurrenceFile corp.cooc -o experiment2 \
-model1iterations 100
```

But this doesn't change which files are produced. How can we see the effect of the change? By asking the program to dump more intermediate results e.g.,

```
./GIZA++ -S eng.vcb -T nor.vcb -C eng_nor.snt \
-CoocurrenceFile corp.cooc -o experiment2 \
-model1iterations 100 -model1dumpfrequency 10
```

Repeat the experiment from the lecture and see that you get the same results after 1,2, 5, 25, 100 iterations. (You may do this by running several experiments with different parameters.)

Compare the results to the output of model3, i.e. with the default settings (5 iterations of model1 followed by 5 iterations of model3).

Part d

Consider the A1 files after the first 5 runs of model1. The results are a little surprising, how? Then consider the result after 100 model1 iterations and after 5 model1 followed by 5 model 3 iterations. What do you see?

Part e

We have made a slightly larger corpus to study effects of alignments which are not one-to-one. You find the texts in /projects/nlp/inf5820/alignment. First of all you should tokenize the texts and also lowercase them. The latter may be done by

```
 /projects/nlp/mosesdecoder/scripts/tokenizer/lowercase.perl \
      < nor.tok > nor.lower
```

Then run GIZA++ in both directions with the default settings. Conider the resulting alignments A3. Draw figures similar to figure 4.13 in the SMT book for the first three sentences.

What to deliver?

- Results from the experiments in part c.
- Answer the questions in part d.
- The figures in part e.

End of alignment