INF5820, Assignment 3: Machine Translation

First part, fall 2012

The assignments for the translation part of INF5820, Fall 2012 will consist in a series of smaller steps where we will experiment with various aspects related to the curriculum and eventually build a translation system with the help of the Moses toolkit. For the translation system we will make use of the tips on the page

`http://www.statmt.org/moses/?n=Moses.Baseline`

Also the manual to which there is a link at that page may be useful. For the first parts, we will try to give detailed explanations—or hints about where to find such info—but the mentioned pages might be of help. What we will do is to look on:

1. Evaluation
2. Word alignment
3. Language model
4. Phrase alignment and system training
5. Tuning

This document describes the first part. The others will be presented in later documents.

1 Evaluation

You will find the relevant material in /projects/nlp/inf5820/evaluation (which you may download to your own file area). There you will find

- 100 sentences of Norwegian (foreign) text: nor100.txt
- Three different reference translations of these sentences translated into English by three different professional translators who are native speakers of English, called ref_b100.txt etc.
- The output of two well known MT systems, lets call them Giggle and Bang in the files sys_x100.txt and sys_y100.txt
- A Bleu script

It is of no importance here which system is Giggle and which is Bang, so please do not test the nor100.txt sentences on the well known web systems.

a The Bleu script works as e.g.

```
./multi-bleu.perl -lc ref_b100.txt < sys_x100.txt
```

(You may also consult the mentioned Moses documentation.) Run this command and see that it works. Take a look at the numbers that you get. This is the Bleu score, followed by unigram, bigram, etc.

The texts are not tokenized. We may get better results if we tokenize and split e.g. “trip.” into two tokens. This may be done with the script

```
 /projects/nlp/mosesdecoder/scripts/tokenizer/tokenizer.perl \
          -l en < ref_b100.txt > ref_b100.tok
```

and similarly for the other texts. Tokenize and check the Bleu score for sys_x100.tok against ref_b100.tok.

b Bleu score sys_y100 against ref_b100 and compare to sys_x100. Then change ref_b100 with ref_c100 and ref_d100 in turn, and check sys_x100 and sys_y100 against each of these. Report the results in a table. Do you find anything surprising about the table?

c Since we do not expect an MT system to do better than a human translator we may take the Bleu score of a human translation as a base line for what to expect from an MT system. Take ref_b100 as the reference and test ref_c100 and ref_d100 as if they were the output of MT systems. Report the results. Are they surprising?

d The Bleu script lets us also compute against several reference translations. To do this we have to rename the reference relations to something on the form ref0, ref1 and eventually ref2 (or file0, file1 etc.; the point is the numbering) and call the Bleu script as

```
./multi-bleu.perl -lc ref < sys_x100.txt
```

Evaluate either system x or y against reference b+c and against b+c+d and compare to evaluating agianst one reference file.

e Let us try to evaluate some sentences manually. Consider the 15 first sentences. Evaluate both translation x and translation y for both adequacy and fluency using a 5 point scale with the same values as on the slide from the lecture. If you don't know Norwegian, use ref_c100 as a reference translation for adequacy. If you know Norwegian, you may use the source text as well. (Strictly speaking, this is not a 100% correct way to do evaluation. Since most of us are not native English speakers we may have problems in judging, and in particular underreport mistakes.)

Give the results in a text file of the form

```
Sentence 1
System X
Adequacy 0
Fluency 0
System Y
Adequacy 0
Fluency 0
Sentence 2
System X
Adequacy 0
...
```

where you exchange the 0s with numbers between 1 and 5.

f Pick 5 translated sentences which you have given the lowest score and explain the shortcomings you find in them.

g Consider the first 15 sentences of reference translation ref_d100 and compare it to the original or with the other reference translations. How do you find this transaltion? Do you think the Bleu score gives a fair impression of its qulity?

End of evaluation