INF5820, fall 2016

Assignment 1: Machine Translation Evaluation

Deadline 23 Sept. at 6 pm, to be delivered in Devilry

1 Examining state-of-the-art MT + error analysis

Since the person who will evaluate your evaluation of the systems only master Norwegian and to some degree English, this part will have two versions depending on which languages you master.

Norwegian speakers. Let English be your source language and Norwegian your target language.

Non-Norwegian speakers. Let English be your target language and choose as source language a language that you know reasonably well.

Choose a recent news text from the web written in your source language and use Google translate to translate it into the target language. Consider (the first) eight sentences, and for each sentence: correct it and report (i) the original sentence (ii) the system output (ii) your manually corrected output (iv) an evaluation of the short-comings of the system's proposal. Also inform about the source of the text.

Here is an example for how this could be done. Text is from dagbladet.no, 5 Sept. 2016: http://www.dagbladet.no/nyheter/i-denne-landsbyen-er-mer-enn-ti-prosent-over-100-ar-gamle-na-avslores-hemmeligheten/62176900

Orig	I denne landsbyen er mer enn ti prosent over 100 år gamle.
Sys	In this village more than ten percent over 100 years old.
Corrected	In this village, more than ten percent of the population are over 100 years old.
Errors	- Lacking copula verb - Lacking noun (*population*). Normal to drop in Norwegian, but not in English.

Orig	Nå avsløres hemmeligheten.
Sys	Now revealed secret.
Corrected	Now the secret is revealed.
Errors	- Passive verb translated as participle without auxiliary. - Wrong word order.

Orig	"Det virker å være en voldsom seksuell aktivitet blant de eldre".
Sys	"It seems to be a violent sexual activity among the elderly.".
Corrected	"It seems to be a vigorous sexual activity among the elderly."
Errors	- "Voldsom" has a literal meaning in Norwegian corresponding to "violent" in English, which can be found in predicative position, like "He is violent". It also has a derived meaning, corresponding to "intense", "energetic", or "vigorous" which prevails when the adjecyive modifies a noun, "voldsom aktivitet".

2 Human MT evaluation

We will try to get a feeling of human evaluation of MT quality. You will find the relevant material in the folder (on the IFI Linux cluster)

`/projects/nlp/inf5820/evaluation`

(which you may download to your own file area). There you will find

- 100 sentences of Norwegian (source) 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, Google and Bing, in the files sys_x100.txt and sys_y100.txt . It is of no importance here which system is which. Please do not test the nor100.txt sentences on the systems, as added material may boost the results on these particular texts in the future.

You shall now evaluate the translation of 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 pure text file (extension .txt) 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. The results shold be delivered exactly in this form, so it will be easy to collect your individual scores. It is also mandatory that you do not collaborate with fellow students on the actual scoring of the sentences, as we will have as many independent assessments as possible.

3 Manual BLEU scoring

A. In this exercise, we shall calculate some BLEU-scores manually. We will consider the translation sys_y100.txt and start with only the first sentence, call it *1y*. As reference translation we will first only use ref_c100.txt. Calculate the BLEU-score for the text consisting of only (1y) against (1c). Ignore punctuation and the difference between capitals and lower case.

B. Use all three reference translations as references and calculate the BLEU-score of (1y).

C. Consider the corpus consisting of the first two sentences, (1y) and (2y). Repeat the two questions for this corpus, i.e., calculate the BLEU score first against ref_c100.txt, and then against all three reference texts.

For all the questions, explain how you calculate the numbers, e.g., show the values for p_1, p_2, etc.

4 Automatic evaluation with BLEU

A. You find a BLEU script which you may download in the folder

`/projects/nlp/inf5820/evaluation`

The script works as e.g.

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

which evaluates the 100 sentences of sys_x100.txt with ref_b100.txt as the only reference translation. 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 tokenizer.perl which you run from where it is by

```
/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. Compare it to the result before tokenization and report the numbers for the two different runs. Answer the question: What do you see?

(You may download the tokenization script to your own area but then you have to include the library it calls.)

B We shall then compare the two MT systems. 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. Answer the questions:

- Which system is best?
- Do you find anything surprising about the table?

C The BLEU script lets us also compute against several reference translations. To do this, we have to rename the reference translations to something of 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 system y against reference b+c and against b+c+d and compare to evaluating against reference b alone. Report in a table.

D Run the BLEU script on the text consisting of the two first sentences of sys_y100 against (the first two sentences of) ref_c100 and compare to your result from the manual calculation.

5 Automatic evaluation of human translation

A We should not expect an MT system to do better than a professional human translator. It might hence be interesting to see how BLEU will evaluate human translation. For this subtask, we take ref_b100 as the only reference translation for calculating the BLEU score. And we will compare the two machine translation systems with the two human translations ref_c100 and ref_d100. Which BLEU score do you get for the two human translations ref_c100 and ref_d100 using ref_b100 as reference, and how do they compare to the two machine translation systems? Report the results in a table. Are they surprising? Explain.

B 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 translation? Do you think the BLEU score gives a fair impression of its quality? Try to speculate why this translation gets such a relatively low BLEU score.

End of evaluation