MPhil in Advanced Computer Science
Module L102: Statistical Machine Translation

Practical Handout 1
Automatic MT Evaluation

1 Introduction

This is the first practical of this module. In total, there will be three practicals. Each practical has its own handout with introduction and questions, which you need to answer in your practical report. Please read carefully this introduction before proceding with the practical. You should prepare a single practical report with all your answers to the questions from all three practicals.

This first practical focuses on automatic evaluation of machine translation using BLEU (Bilingual Evaluation Understudy), the most widespread evaluation metric within the research community. You will be asked to compute BLEU scores for a set of translations from Arabic into English. For this purpose, a freely-available scoring script is provided. Run the following command to have a direct path to this script:

```
export PATH=$PATH:/usr/groups/acs-software/L102/practical-1/bin
```

1.1 Setting

We will evaluate the quality of multiple Arabic→English translation systems on a 10-sentence set. The 10-sentence input Arabic text, 4 English reference translations (created by four independent human translators) and 2 automatic translations are provided in the following files:

Input Arabic text:	`$DIR/input/10lnsv1.ara`
English reference 1:	`$DIR/reference/10lnsv1.1.eng`
English reference 2:	`$DIR/reference/10lnsv1.2.eng`
English reference 3:	`$DIR/reference/10lnsv1.3.eng`
English reference 4:	`$DIR/reference/10lnsv1.4.eng`
Output System 1:	`$DIR/output/10lnsv1.cued_v1.eng`
Output System 2:	`$DIR/output/10lnsv1.cued_v2.eng`

where `DIR=/usr/groups/acs-software/L102/practical-1`

1.2 Example

In order to compute the BLEU score of a set of translations against one or more references you will need to use `ScoreBLEU.sh`:

```
Usage: ScoreBLEU.sh -t hyp -r ref1 [-r ref2... -r refN]
  -t       : translation hypothesis to be evaluated
```

```
  -r       : one (or more) reference translation/s
optional:
  -d       : detailed output
  -odir    : output directory (default: scoring/ )
  -case    : preserve case (by default, case insensitive)
```

For example, the following command:

```
ScoreBLEU.sh -t $DIR/output/10lnsv1.cued_v1.eng
   -r $DIR/reference/10lnsv1.1.eng
```

returns the BLEU score obtained by System 1 when comparing against the first English reference. Check that this gives:

```
BLEU score = 0.2822 (0.2822 * 1.000) for system "1"
```

which shows a percentual BLEU score of 0.2822, as obtained by multiplying the cumulative 4-gram precision by the brevity penalty (shown in brackets). Typically this is reported on a percentual range, that is 28.22.

The scoring script actually outputs some extra information, including sentence-level BLEU scores and individual and cumulative n-gram precisions for n from 1 to 9. This is written into a file in the output directory selected with the `-odir` (by default: `./scoring` is used), but can also be shown on the standard output by using the `-d` switch.

Check that the sentence-level BLEU for sentences 3 and 4 are 0.4312 and 0.1312, respectively.

2 Practical Exercise

1. Review the lecture notes and write down the BLEU score evaluation metric formula and explain how it is computed, and the role of the brevity penalty.

   (a) Is the final score a linear combination of the BLEU score obtained for each sentence? If we are scoring a set of two sentences of 10 and 25 words in length each, which one will have a stronger impact on BLEU? Why?

   (b) What is different between computing BLEU against one single reference translation and against a set of multiple references?

2. Fill in the following Table where each of the two sets of automatic translations are to be evaluated against each of the four references, and against multiple references. Some values have already been computed for you. Please check that you also obtain the same scores for these cases. Note that the brevity penalty is also to be reported.

	Reference Translation/s					
System	r1	r2	r3	r4	r1,2	r1,2,3,4
cued_v1	28.2 (1.000)		25.3 (0.891)	27.6 (0.971)	35.6 (1.000)	
cued_v2		22.3 (1.000)				

(a) Discuss the differences in scores when using each of the individual references. By inspecting the brevity penalty in each case, determine which human translator produced the longest reference. Corroborate your answer by counting the words in each reference file (simply use '`wc -w`').

(b) Discuss the differences in scores when using one, two or four English references. Is the BLEU difference between both systems constant? Discuss also any changes in brevity penalty.

(c) In view of your previous answer, which of the 6 columns provides the best comparison between systems? Which of the two systems is yielding a better translation for these 10 sentences?

3. Let us now evaluate the agreement between human translators. In order to do so, compute the BLEU score for each of the four English reference translations against the remaining three.

   (a) Which of the four references is the most different with respect to others? Which is the most similar?

   (b) Now discard the most different reference and re-compute BLEU scores for the two systems with respect to the remaining 3 references. Compare these results with the score the most different reference obtained against the same three references. According to this, can we conclude that the automatic systems produce a better translation than a human professional? Why, or why not?

4. Translate the 10 sentences using at least 3 free online MT services:

   - Reverso (`http://www.reverso.net`)
   - Google Translate (`http://translate.google.com`)
   - Bing (`http://www.microsofttranslator.com/`).

   Make sure you copy and paste your text properly into these websites, as it may be necessary to translate in two/three batches if there is a limitation on the character length. Please report what date the translation was done. If you use any additional service, please report which. Extend the table of Question 1 with one row per new system.

	Reference Translation/s					
System	r1	r2	r3	r4	r1,2	r1,2,3,4
cued_v1	28.2 (1.000)		25.3 (0.891)	27.6 (0.971)	35.6 (1.000)	
cued_v2		22.3 (1.000)				
reverso						
google						
bing						
...						

5. Discuss the ranking of systems when evaluated against each of the references, or against 2 or 4 references. Is it always the same? Which system(s) performs best and which system performs worst according to these results? What do you think is the most reliable ranking?

6. Suppose we would like to include a fifth English reference translation to have a more reliable evaluation scheme. Of course, one way of creating this would be to pay yet another professional to translate the Arabic input into English. However, a cheaper option could be to take one set of automatic translations and simply carry out the minimal manual corrections required in order to turn the output sentence into natural English while retaining the same meaning as other references.

   (a) By examining the first sentence from `$DIR/output/10lnsv1.cued_v1.eng` and the first sentence of each of the references, manually correct the output sentence with the minimum possible changes. Write down your corrected sentence in your report.

   (b) The same manual correction has been done for you on the remaining 9 sentences in the following file:

   `$DIR/output/correction/10lnsv1.cued_v1.correct.eng`

   Change the first line of this file by your corrected output to create a 5th reference file (`10lnsv1.5.eng`). Score all systems against the 5 references. Is the ranking the same as when using the 4 references? What ranking is more reliable now?

   (c) What is the problem with the previous contrast with 5 references? What would be a fairer way of generating new references?