Machine Translation: Practical Work 1

Sharid Loáiciga

Set up

1. Get the evaluation tools MultEval (Clark et al. 2011):

   `wget http://www.cs.cmu.edu/~jhclark/downloads/multeval-0.5.1.tgz`

2. Unpack them:

   `tar -xvzf multeval-0.5.1.tgz`

3. Run the script. This will call Meteor that is most probably not currently installed in your computer yet. Therefore the command will fetch and install it. This step can take 10-15 minutes. (Note that the following two lines are to be entered as a single command, as indicated by the symbol \)

   `./multeval.sh eval -refs example/refs.test2010.lc.tok.en.* \`

   `-hyps-baseline example/hyps.lc.tok.en.baseline.opt* -meteor.language en`

1 Crete multi-reference datasets (online)

Pair up with a colleague who knows the same pair of languages a you do.

We have prepared three datasets that contain sentences in English, be we do not have any reference translation for these files yet. Your first task is to help us create multi-reference datasets by translating these sentences to another language. We will use the resulting datasets in exercise 3.

Decide which file you will translate into which language. You will do the translations online. Please try to create the translations without using online MT tools.

- Set A `https://forms.gle/3WiHc3nd4BDMuWjn6`
- Set B `https://forms.gle/Mjv3hX4EieVjjoL6A`
- Set C `https://forms.gle/cgEqWFgSvyxRMuqS8`

2 Manual and automatic evaluation

Unpack the folder `exercise2.zip` from `https://github.com/compling-potsdam/sose20-mt/tree/master/docs/materials/pw1/`. The file `manual-evaluation.txt` contains 20 sentences that have been translated from Finnish to English using four different translation systems. For each sentence, a reference translation is also given.

1. Propose a ranking for each sentence. Consider adequacy as well as fluency when proposing the ranking. For example, if system 4 is best, system 3 is second-best, and system 1 is worst, the ranking would look like this: 4 3 2 1. Write the rankings on the appropriate lines of the `manual-evaluation.txt` file.

2. In the same folder, there is a Python script `averageRanking.py` that will compute the average ranks of the systems. Run the script as follows and record the result:

   `python3 averageRanking.py`

3. There is at least one rule-based MT system, one statistical system and one neural system among the four. Can you guess which one is which? Justify your hypotheses.

4. The folder contains the same sentences in a different file format, with one file per system. Compute the BLEU scores all the systems with respect to the reference translation using the following command:

   `./multi-bleu-detok.perl reference1.txt < system1.txt`

   How well do they correlate with the average ranks obtained from your personal judgements?

   The multi-bleu-detok.perl script also accepts a parameter -lc which converts all system outputs and references to lower case. How does this affect the BLEU scores? Does it change the ranking of the four systems?

   `./multi-bleu-detok.perl -lc reference1.txt < system1.txt`

3 Single-reference and multi-reference evaluation

Unpack the folder `exercise3.zip` available here: `https://github.com/compling-potsdam/sose20-mt/tree/master/docs/materials/pw1/`. It contains the multi-reference datasets created in exercise 1, separated according to target language. Choose one dataset to work with (it does not have to be the same one that you have translated). With each dataset, we also provided a `system1.txt` file, which corresponds to a machine translation output.

Compute automatic evaluation scores comparing single-reference and multi-references scenarios.

You could use the `multi-bleu-detok.perl` script (as in exercise 2) for this, but the `multeval-0.5.1` package installed at the beginning computes BLEU as well as TER and METEOR.

For instance, scenario with one Spanish reference:

`$ ./multeval-0.5.1:` $\rightarrow$ This means "from within the folder /multeval-0.5.1:"

```
./multeval.sh eval \
-refs reference1.txt\
-hyps-baseline system1.txt \
-meteor.language es
```

... with several Spanish references :

```
./multeval.sh eval \
-refs reference2.txt reference1.txt \
-hyps-baseline system1.txt \
-meteor.language es
```

4 Submission

Submit a PDF file in Moodle by Wednesday March 6th, 23:00. You should document your progress in the assignment. Note your observations and findings as well as the output of the evaluation scripts. For exercise 1, just state which set and language was assigned to you. Together with the PDF, please also hand in the file with your annotated rankings of exercise 2.

References

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith (June 2011). “Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 176–181. URL: https://www.aclweb.org/anthology/P11-2031.