Machine Translation Quality and Error Analysis

by Maksym Del & Mark Fishel

The next adventure on your shared task journey is to take a closer look at the quality of todays's neural machine translation systems.

Translate in-domain dev set with you baseline system

1. Pull lab3 folder from the materials repository
2. Read steps 4 and 5
3. Translate the in-domain dev set (*accurate*) from the lab3 folder with your baseline NMT system and report BLEU score. Compare it with your in-domain dev-set and conclude on difference.
4. Do not forget to DeBPE, detokenize and detruecase the dev set before doing step 2 (since you claimed to do so in your baseline reports ☺)
5. Do not forget to translate the dev set using OpenNMT-py from the lab3 folder, which outputs attentions to use with attention viz. tool we studied in lab. Passing *-attn_debug* parameter to the *translate.py* leads to outputting both *hyps* and *attentions* files.

Analyse the Translations

Write a report about the quality of the machine translation.

Go over at least 40 sentences (60 for 3-person teams), manually inspect each sentence, and report for each sentence:

1. sentence id (line number)
2. the source sentence
3. reference translation
4. the machine translation
5. an assessment of the error in the machine translation

You may do step 5 in any way you want. For instance, you could classify errors as "reordering errors", "untranslated words in source", or any other type of error you can think of.

For instance:

1. ID: 1984

2. *Erst drei Tage ist der neue Ministerpräsident Griechenlands im Amt.*
3. *The new Prime Minister of Greece has been in office for only three days*
4. *Only three days is the new Prime Minister of Greece in office.*
5. (1) Verb tense is wrong: *is* instead of *has been*. (2) Preposition *for* was missing in front of the time phrase *only three days*. (3) While the noun phrases and preprositional phrases are correct, the overall sentence structure on the clause level is scambled.

   (adapted from http://mt-class.org/jhu-2016/hw0.html):

Another example:

1. ID: 2019
2. *kirjalikult . - (IT) Austatud juhataja , daamid ja härrad . mina hääletasin härra Gyürki riiklikke energiasäästu tegevuskavasid järelkajastava raporti poolt .*
3. *in writing . - (IT) Madam President , ladies and gentlemen , I voted in favour of Mr Gyürk 's report on follow-up of the energy efficiency national action plans .*
4. *in writing . - (IT) Madam President , ladies and gentlemen , I voted in favour of Mr Beaupuy 's report on the national energy saving action plans .*
5. Good translation with perfect fluency. (1). But minor adequacy problems: guessing the *"Madam" President* from the genderless Estonian input wrongly (2). The name *"Gyürk"* is translated as *"Beaupuy"* (segmented as *Gy-ür-ki* by BPE) (3)

Also take a loot at https://github.com/fishel/kama/ (open nmt and smt subfolders).

Conclude your report with a summary of your impression of the major quality problems in the machine translation system that you analysed.

Analyse attentions

Take a look at metrics from https://ufal.mff.cuni.cz/pbml/109/art-rikters-fishel-bojar.pdf (section 4.1. Scoring Attention). Feed your 40 (60) sentences into the system and sort from worst to best by the metrics from the paper. Add screenshot to the report. Is there any correspondence between metric and the most problematic sentences you found? Answer shortly in you report.

Push your report to github (md or pdf) before milestone deadline and do not forget to add tasks to the Project Board and your Project Board screenshot to the report.