BBC News Labs – Coding Exercise

Background

One of our current projects, called GoURMET, aims to develop high-quality machine translations for languages where very little training data exists. BBC News Labs would like to use these to build new tools for our journalists and we are currently assessing the quality of some of the machine translation models that have been developed within the project.

The BBC publishes content in over 40 languages around the world. Our machine translations could enable us to more quickly “reversion” our content, creating separate versions for different languages and reversioning is a key part of efficiently reporting on world events.

One way to test if a translation model is suitable for reversioning content is using an evaluation technique called Direct Assessment. In a Direct Assessment test a native speaker of the language being evaluated is shown a sentence translated by a human and the same sentence translated by the translation model. The speaker assesses on a scale from 0 to 100 whether the machine translation “adequately expresses” the meaning of the human translation.

For example, to test the quality of a Bulgarian model we take an English sentence and get a bilingual human to translate it into Bulgarian. This is sentence 1. The same sentence is also translated using the machine translation model. This is sentence 2. A native Bulgarian speaker is then asked if sentence 2 “adequately expresses” the meaning of sentence 1 on a scale from 0 to 100.

In the GoURMET project, we have been using Direct Assessment to evaluate the quality of a translation model produced by our academic partners to translate from English into Bulgarian.

Dataset

In this exercise, we provide you with a subset of this data collected during Direct Assessment evaluation. Each row shows the 2 translations and the original text along with the score that an evaluator has given. Each sentence has been scored multiple times by different evaluators.

File[1]:

- bulgarian-direct-assessment.csv

Each row is made up of:

- **`sentence id`**: a unique id for each sentence (the same sentence will have been evaluated multiple times)
- **`evaluator id`**: a unique id for each evaluator (the same evaluator will have evaluated multiple sentences)
- **`score`**: the score the evaluator gave each machine translated sentence
- **`human translation`**: the sentence as translated by a human
- **`machine translation`**: the sentence as translated by a machine
- **`original`**: the original sentence in English

Summary

For this exercise, we'd like to ask you to write some code to read in the data and generate outputs for the exercises below.

Use the programming or scripting language and an approach with which you feel most comfortable (Python/Javascript/R/other languages, and language tools, constructs and libraries etc.)

We suggest that you should aim to spend *no more than 2-3 hours* on the exercise.

Please submit, via an email to `coding-exercises@bbcnewslabs.co.uk`, *no later than* **Tuesday 4th February**, either a link to a github repository or a zip file containing:

- Your code
- Your outputs
- A README file with instructions on how to install/run your code

Please feel free to include a text file containing a brief outline (no more than a paragraph or two) of any choices you have made.

---

[1] Opening CSV in Microsoft Excel:
The CSV contains Cyrillic characters which do not always render correctly when the file is opened in Microsoft Excel. To view the CSV in Excel:
1. Open a blank workbook in Excel
2. Select the "Data" tab
3. Select "From Text"
4. Select the CSV file then "Get Data"
5. The "Delimited" option should be selected and the file origin should be "UTF-8" then select next
6. Choose "Comma" as the delimiter type
7. Click finish

Exercise

For each exercise start with the original CSV file

1. Output the average score given by each evaluator
2. Output the average score for each sentence
3. Output the highest and lowest scoring sentences.

During the interview, we'd like to spend a few minutes discussing your solution and the approach you've taken.

If you have any questions or need any clarifications relating to this exercise, please email `coding-exercises@bbcnewslabs.co.uk`.