

Text Mining Project

Nova IMS, 2021

NOVA IMS MT Metrics Shared Task

The conference on Statistical Machine Translation (WMT) is a prestigious conference in the field of NLP and Machine Translations that builds on a series of annual workshops and conferences in Machine Translation, going back to 2006. This prestigious conference organizes a series of competitions in different aspects of machine translation. One of those competitions is the **Metrics shared task** whose goal is to create a metric that correlates with human assessments of quality.

Metrics shape the direction of research. They allow us to compare experiments, understand what our models are doing wrong, identify problems to work on, etc. Last year, I participated in this prestigious task and this year I'll be working as an organiser. This experience has been very rewarding for me and for that reason I decided that **during this course you will have the pleasure to participate in a simulated version of this shared task.**

With this said, I will release a corpus with translations for the following language pairs:

- Russian into English
- German into English
- Czech into English
- Chinese into English
- English into Chinese
- English into Finish

For each translation you have the original segment, a reference translation (you gold truth) and quality assessment reflecting "how good is a translation" in a scale of 0 to 100, given by 1 or more human annotators. Those scores are then normalised to produce the z-score.

Example of a bad translation in our corpus:

Original text: *Diese Fähigkeit sei möglicherweise angeboren oder entwickle sich mit der Geschlechtsreife.*

Translation: *This ability may be innate, or may develop as the animal reaches sexual maturity.*

Reference: *This ability may be born or developed with gender maturity.*

Quality assessment according to two annotators: 51.5

Normalised score: -1.2565720794229098

Example of a good translation in our corpus:

Original text: *Er sagte, dass die Bereiche ruhige Treffpunkte zwischen Flüchtlingen und Freiwilligen bieten.*

Translation: *He said the spaces provided calm meeting points between refugees and volunteers.*

Reference: *He said the areas offer quiet meeting points between refugees and volunteers.*

Quality assessment according to two annotators: 97.5

Normalised score: 0.903800434358589

The goal of this project is to develop a metric that predicts the quality of a translation using the reference. Your metric should correlate well with the existing quality assessments that you have in the above corpus.

Evaluation:

Closer to the deadline I'll release a test set with CSV's without the columns z-score, avg-score and annotators. For each CSV file you will have to add a column with name "metric" containing the scores produced by your metric. Then you will zip that test set and submit it through Moodle. We will open those files and compare the outputs from your metric with the ground-truth scores using Pearson and Kendall Tau. Your results will be made available in a final leaderboard.

We also expect a report with 3 content pages (with unlimited appendix and bibliography) describing your solution and where you report correlations between your metrics and a development set developed by you.

Important Notes:

- The delivery deadline for the project and report is May 28th, at 17:00.
- Students should not use code from each other!
- The project is to be developed in groups of up to three.

Comments/advises:

A Machine Translation metric can be something simple that simply measures the lexical similarity between a translation and a reference or it can be a complex deep learning system that tries to directly estimate the human assessments. If you really feel inspired you can try to leverage information from the source segments and/or try to solve the task without any reference!

Academic Honesty / Honor Code

Any instance of cheating diminishes the NOVA brand and is a violation of other students right to fairness and justice. Instances of cheating include:

- Plagiarism - NOVA has a strict policy against the deliberate reproducing of work of another person or institution without acknowledgement. All sources used for any piece of work should be fully referenced and acknowledged.
- Unlawful Collusion - Collusion between students in the production of materials included in the grading process, whether taking place in the classroom or at home, other than teamwork explicitly assigned by the teacher in unlawful is not allowed. Students are expected to complete their own work.
- Unlawful Copying - Copying from another student's examination, with or without his/her consent, using course material during a closed book exam, submitting the same work for more than one course without prior permission of both instructors, violating any examination recommendations or any rules relating to academic conduct of a course, unauthorised use of cell phones, calculators, dictionary, books, computer during an examination, are instances of unlawful copying.