Evaluating Machine Translation as preprocessing step

Laboratory:

Swiss Data Science Center

Type:

Semester Project

Description:

Natural Language Processing has made recent progress partly due to the introduction of the Transformers architecture and large scale datasets sourced from the web. However, as English makes up more than 60% of the World Wide Web, most of the models suffer performance drop when switching to other languages.

Previous works [1, 2] have shown that it is possible to use Machine Translation (MT) to process non-English text into English and then apply SOTA models trained on English data. However, these studies focus on sentiment analysis and ignore other downstream tasks such as text classification or named entity recognition.

In this project we aim to do an exhaustive evaluation of the trade-off between the use of MT models to perform pre-processing on non-English and the use of specialized language models.

Goals/benefits:

- Working with machine learning and deep learning libraries in Python (pandas, scikit-learn, PyTorch)
- Becoming familiar with state-of-the-art NLP models
- Advancing research on an interdisciplinary problem

Prerequisites:

- Machine learning and deep learning (advanced or intermediate skills)
- Python (advanced skills)

Deliverables:

- Well-documented code
- Written report and oral presentation

References:

[1] An evaluation of machine translation for multilingual sentence-level sentiment analysis (Araujo et al., Proceedings of the 31st Annual ACM Symposium on Applied Computing 2016)

[2] Machine Translation for Machines: the Sentiment Classification Use Case (Tebbifakhr et al., EMNLP 2019)

Contact:

Lefebvre Clément (clement.lefebvre@epfl.ch)