Task 4: Machine translation

Data

FLORES dataset
https://github.com/facebookresearch/flores/tree/main/flores200
Goyal, N., Gao, C., Chaudhary, V., Chen, P. J., Wenzek, G., Ju, D., ... & Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10, 522-538.

Huggingface

HF Translation task: https://huggingface.co/tasks/translation

HF Course: https://huggingface.co/learn/nlp-course/chapter7/4

Subtasks and points

1. Read the HF documentation. Find HF translation models that support English-Kazakh (alternatively English—Russian) translation.
2. Start with Facebook's NLLB model https://huggingface.co/facebook/nllb-200-distilled-600M Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
3. Read the description of the BLEU score https://en.wikipedia.org/wiki/BLEU and its implementation https://github.com/mjpost/sacrebleu
4. Evaluate the NLLB model using the *devtest* subset of the FLORES dataset (evaluate both directions: En—Xx and Xx—En). (20)
5. Collect a small (~50 sentences) parallel corpus for your language pair. Use two different genres, e.g. news and fiction. Align the data; evaluate the NNLB model on the dataset. (20)
6. Evaluate another model from the HF on two datasets (FOLORES and your own). (20)
7. Evaluate Google Translate and Yandex Translate on the same datasets. (20)
8. Summarize and analyze the evaluation results (BLEU scores). Manually inspect sentence pairs with lowest/highest scores and provide analysis. (20)