MUNI FI

Machine Translation

PV061

Pavel Rychlý

NLP Centre, FI MU

20 Sep 2023

BPE

Subword Neural Machine Translation

- https://github.com/rsennrich/subword-nmt
- `pip install subword-nmt`

SentencePiece

- https://github.com/google/sentencepiece
- `pip install sentencepiece`
- python wrapper:
  https://github.com/google/sentencepiece/blob/master/python/README

Shared vocabulary

- encoder-decoder
  - each part separate word embeddings
  - decoder: separate input/output embeddings

Fairseq

- https://github.com/facebookresearch/fairseq
- MT example:
  https://github.com/facebookresearch/fairseq/tree/main/examples/trans
- checkpoint is a single file

HuggingFace

- https://huggingface.co/
- `pip install transformers`
- https://github.com/huggingface/transformers/tree/main/examples/pyt
- pretrained models
- datasets: `export HF_DATASETS_CACHE=/big-disk/datasets`

MUNI
FACULTY
OF INFORMATICS