11-737 Multilingual NLP: Assignment 2

Patrick Fernandes & Vijay Viswanathan

Carnegie Mellon University
Language Technologies Institute

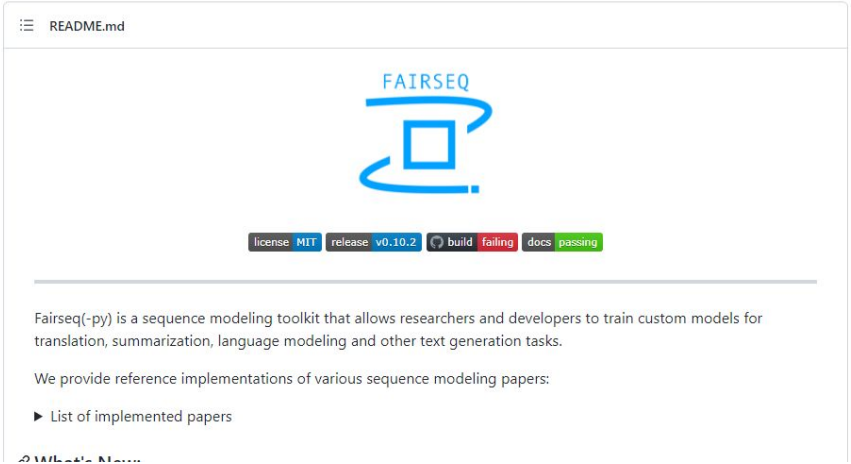
Task: Machine Translation

- ➔ Given a *source-language* sentence, you want to translate it to the *target language*
- ➔ The goals of this assignment are:
  - ◆ Understand the standard data preprocessing pipeline used for MT
  - ◆ Be able to train MT models, both bilingual and multilingual, using an MT framework
  - ◆ Learn how these models are evaluated
  - ◆ Investigate methods to tackle the *data-scarcity* problem in low-resource language pairs

Requirements

- ➔ You NEED a machine with a GPU
- ➔ Same requirements as last assignment
  - ◆ Conda
  - ◆ Python=3.8
  - ◆ PyTorch=1.10.1
- ➔ This assignment will also use *fairseq* as the backbone for training models

Fairseq



➔ Automates data-loading, training and decoding

➔ Supports many tasks other than translation

Requirements

- ➔ You NEED a machine with a GPU
- ➔ Same requirements as last assignment
  - ◆ Conda
  - ◆ Python=3.8
  - ◆ PyTorch=1.10
- ➔ This assignment will also use *fairseq* as the backbone for training models
- ➔ You also need *sacreBLEU* and *COMET* to evaluate your models

assign2.zip

```
assign2
├── download_data.py
├── …
├── preprocess-ted-{bilingual,multilingual,flores-bilingual}.sh
├── traineval_{aze,eng}-{eng,aze}.sh
├── traineval_{azetur,eng}-{eng,azetur}.sh
├── traineval_flores_{aze,eng}-{eng,aze}.sh
└── score.py
```

Data

- ➔ You will be using the *TED talks* [1] corpus
  - ◆ Contains parallel data between english and 58 languages
- ➔ You will focus on two *low-resource* language pairs
  - ◆ English-Azerbaijani
  - ◆ English-Belarusian

https://github.com/neulab/word-embeddings-for-nmt

Preprocessing

- ➔ Consists of following steps different
  - ◆ Read raw parallel data
  - ◆ Learn byte-pair encoding (BPE) separately for source and target language(s) on train set
    - In this assignment, SentencePiece is used
  - ◆ Apply BPE to all splits
  - ◆ Do some very simple data cleaning to the training set
  - ◆ Binarize data

Modeling

- ➔ Transformer architecture
- ➔ Embeddings are shared between source and target
- ➔ Trained to minimize Cross Entropy with Adam
- ➔ Decoding is done using beam search

```
fairseq-train \
        $BINARIZED_DATA \
        --task translation \
        --arch transformer_iwslt_de_en \
        --max-epoch 80 \
        --patience 5 \
        --distributed-world-size 1 \
        --share-all-embeddings \
        --no-epoch-checkpoints \
        --dropout 0.3 \
        --optimizer 'adam' --adam-betas '(0.9, 0.98)' --lr-scheduler 'inverse_sqrt' \
        --warmup-init-lr 1e-7 --warmup-updates 4000 --lr 2e-4  \
        --criterion 'label_smoothed_cross_entropy' --label-smoothing 0.1 \
        --max-tokens 4500 \
        --update-freq 2 \
        --seed 2 \
        --save-dir $MODEL_DIR \
        --log-interval 20 2>&1 | tee $MODEL_DIR/train.log
```

```
fairseq-generate $BINARIZED_DATA \
    --gen-subset test \
    --path $MODEL_DIR/checkpoint_best.pt \
    --batch-size 32 \
    --remove-bpe sentencepiece \
    --beam 5  | grep ^H | cut -c 3- | sort -n | cut -f3- > "$MODEL_DIR"/test_b5.pred
```

Evaluation

- ➔ Evaluation is based on two automatic evaluation metrics:
  - ◆ **BLEU**
    - Uses n-gram overlap between reference and target
  - ◆ **COMET**
    - Uses neural encodings of reference and target

Bilingual Training

➔ You will start by training a model just on the parallel corpora for the *low-resource* language pairs

```
preprocess-ted-bilingual.sh

traineval_{aze,eng}-{eng,aze}.sh
```

➔ You will need to modify these scripts slightly for Belarusian

Multilingual Training

- ➔ Your bilingual models will have a very low performance
- ➔ To improve it, you will experiment with doing multilingual training, transfering from a *high-resource* LP
  - ◆ For Azerbaijani, you will start with Turkish

```
preprocess-ted-multilingual.sh

traineval_{azetur,eng}-{eng,azetur}.sh
```

Fine-tuning Massive Multilingual models

- ➔ An alternative way to improve models is to finetune a massive multilingual model on this data
- ➔ In this assignment we will consider fine-tuning the FLORES-101 (small) model
- ➔ You will need to download the checkpoint files (instructions in the homework)

```
preprocess-ted-flores-bilingual.sh
traineval_flores_{aze,eng}-{eng,aze}.sh
```

Extensions

- ➔ Data Augmentation
  - ◆ Back-translation
  - ◆ Self-training
  - ◆ Many others

Extensions

- ➔ Better Transfer Languages
  - ◆ Apply your knowledge of typology to improve performance
  - ◆ Automatically choosing transfer languages

Extensions

- ➔ Better Word Segmentation
  - ◆ Does something other than BPE work better?
  - ◆ Other techniques for open-vocabulary MT (such as visual text representations)?

Extensions

- ➔ Better modeling and training choices
  - ◆ Be creative!

Submission

- ➔ Three deliverables
  - ◆ Code
  - ◆ Writeup
  - ◆ Model Outputs
    - to reproduce your evaluations
- ➔ Submitted as a tarball on Canvas

Grading

- **B:** Reproduce the bilingual results and either the multilingual training results OR finetuning results
  - For both Azerbaijani and Belarusian

Grading

- ➔ **B:** Reproduce the bilingual results and either the multilingual training results OR finetuning results
  - ◆ For both Azerbaijani and Belarusian
- ➔ **B+:** Detailed analysis of how multilingual training/finetuning improves performance
  - ◆ What type of source sentences especially benefit from multilinguality?

Grading

- ➔ **B:** Reproduce the bilingual results and either the multilingual training results OR finetuning results
  - ◆ For both Azerbaijani and Belarusian
- ➔ **B+:** Detailed analysis of how multilingual training/finetuning improves performance
  - ◆ What type of source sentences especially benefit from multilinguality?
- ➔ **A-:** Implement at least one pre-existing method to try to improve multilingual transfer
  - ◆ Compare with the provided methods and analyse results

Grading

- **B:** Reproduce the bilingual results and either the multilingual training results OR finetuning results
  - For both Azerbaijani and Belarusian
- **B+:** Detailed analysis of how multilingual training/finetuning improves performance
  - What type of source sentences especially benefit from multilinguality?
- **A-:** Implement at least one pre-existing method to try to improve multilingual transfer
  - Compare with the provided methods and analyse results
- **A:** Implement several methods to improve multilingual transfer:
  - Multiple pre-existing methods or one pre-existing and one novel method

Grading

- **B:** Reproduce the bilingual results and either the multilingual training results OR finetuning results
  - For both Azerbaijani and Belarusian
- **B+:** Detailed analysis of how multilingual training/finetuning improves performance
  - What type of source sentences especially benefit from multilinguality?
- **A-:** Implement at least one pre-existing method to try to improve multilingual transfer
  - Compare with the provided methods and analyse results
- **A/A+:** Implement several methods to improve multilingual transfer:
  - Multiple pre-existing methods or one pre-existing and one novel method
  - Particularly extensive and/or interesting analysis/methods will earn a **A+**

Additional Help

- Office Hours (check piazza for location/zoom links):
  - Patrick - Tuesday 3:30-4:30PM
  - Vijay - Monday 3-4pm

- Contact
  - Patrick: pfernand@cs.cmu.edu
  - Vijay: vijayv@andrew.cmu.edu
  - TA Mailing List: cs11-737-sp2022-tas@cs.cmu.edu