Homework 2
Ling 83600 Language Technology
Tawa Suleman
27 October 2021

Part 2: ***Preprocessing***

This part went fairly smoothly. However, I noticed that the .p files had extra lines between the transcriptions, so I had to edit my script to address and fix this. Below is the logging output from when I ran fairseq-preprocess:

```
2021-10-15 23:18:00 | INFO | fairseq_cli.preprocess | Namespace(align_suffix=None, alignfile=None, all_gather_list_size=16384, bf16=False, bpe=None, checkpoint_shard_count=1, checkpoint_suffix='', cpu=False, criterion='cross_entropy', dataset_impl='mmap', destdir='data-bin', empty_cache_freq=0, fp16=False, fp16_init_scale=128, fp16_no_flatten_grads=False, fp16_scale_tolerance=0.0, fp16_scale_window=None, joined_dictionary=False, log_format=None, log_interval=100, lr_scheduler='fixed', memory_efficient_bf16=False, memory_efficient_fp16=False, min_loss_scale=0.0001, model_parallel_size=1, no_progress_bar=False, nwordssrc=-1, nwordstgt=-1, only_source=False, optimizer=None, padding_factor=8, profile=False, quantization_config_path=None, scoring='bleu', seed=1, source_lang='ice.g', srcdict=None, target_lang='ice.p', task='translation', tensorboard_logdir=None, testpref='test', tgtdict=None, threshold_loss_scale=None, thresholdsrc=2, thresholdtgt=2, tokenizer='space', tpu=False, trainpref='train', user_dir=None, validpref='dev', workers=1)
2021-10-15 23:18:00 | INFO | fairseq_cli.preprocess | [ice.g] Dictionary: 72 types
2021-10-15 23:18:01 | INFO | fairseq_cli.preprocess | [ice.g] train.ice.g: 800 sents, 5045 tokens, 0.872% replaced by <unk>
2021-10-15 23:18:01 | INFO | fairseq_cli.preprocess | [ice.g] Dictionary: 72 types
2021-10-15 23:18:01 | INFO | fairseq_cli.preprocess | [ice.g] dev.ice.g: 100 sents, 609 tokens, 1.48% replaced by <unk>
2021-10-15 23:18:01 | INFO | fairseq_cli.preprocess | [ice.g] Dictionary: 72 types
2021-10-15 23:18:01 | INFO | fairseq_cli.preprocess | [ice.g] test.ice.g: 100 sents, 643 tokens, 1.24% replaced by <unk>
2021-10-15 23:18:01 | INFO | fairseq_cli.preprocess | [ice.p] Dictionary: 64 types
2021-10-15 23:18:01 | INFO | fairseq_cli.preprocess | [ice.p] train.ice.p: 1600 sents, 6176 tokens, 0.0486% replaced by <unk>
2021-10-15 23:18:01 | INFO | fairseq_cli.preprocess | [ice.p] Dictionary: 64 types
2021-10-15 23:18:02 | INFO | fairseq_cli.preprocess | [ice.p] dev.ice.p: 200 sents, 752 tokens, 0.133% replaced by <unk>
2021-10-15 23:18:02 | INFO | fairseq_cli.preprocess | [ice.p] Dictionary: 64 types
2021-10-15 23:18:02 | INFO | fairseq_cli.preprocess | [ice.p] test.ice.p: 200 sents, 785 tokens, 0.255% replaced by <unk>
2021-10-15 23:18:02 | INFO | fairseq_cli.preprocess | Wrote preprocessed data to data-bin
```

Part 3: ***Training***

I was able to find most parameters for our model through the provided fairseq docs. I was not able to find the appropriate flag for setting the encoder as bidirectional, and for setting the smoohing coefficient. Conferring with others I was able to find the flags I needed. One pair of flags I had used were incorrect, which caused my training to run for an unnecessarily long time. I had initially set the layers to 512 (eg. --encoder-layers 512) instead of the hidden layer size. Once this was fixed, my training went much more smoothly. Below is the fairseq-train command I ran:

```
fairseq-train data-bin --source-lang ice.g --target-lang ice.p --seed 925 --arch lstm --encoder-bidirectional  --dropout .2 --encoder-embed-dim 128 --decoder-embed-dim 128 --decoder-out-embed-dim  128  --encoder-hidden-size  512  --decoder-hidden-size  512  --criterion  label_smoothed_cross_entropy  --label-smoothing  0.1  --optimizer  adam  --lr 0.001 --clip-norm 1 --batch-size 50 --max-update 800 --no-epoch-checkpoints
```

Part 4: ***Evaluation***

The main issue I came across when evaluating predictions.txt was correctly retrieving the target and predictions from the file. On my initial attempt I had incorrectly chosen the indices and would not take in the entire words. Once this was fixed, though, I was able to evaluate the predictions and find a WER of 28.