FLoRes baseline reproduction

Below are the results for my FLoRes reproduction experiments. `NE -> EN` was completed on Microsoft Azure, and the rest on Amazon EC2. In all experiments, the GPU used was a Tesla K80. Overall, I ran everything for 100 epochs, with about 20 min per epoch being the average runtime.

Correspondence with FAIR team

As you know, I submitted an issue on the `flores` GitHub page, describing my setup in detail. It turns out that they actually *did* respond! Here are some things I found out:

- Very confusingly, the `test` set in the training/eval scripts is indeed what corresponds to `devtest` in the paper. They added a note to he README after I complained about this.
- The `fairseq` version they used was 7.2
- Regarding randomness in `fairseq`, apparently it shouldn't matter as they have *"handled the seed carefully in our released pipeline, including both data preprocessing and training pipeline (including fairseq itself)"*.
- Apparently there *is* slight variation that is to be expected when reproducing the experiments: *"In multi-GPU cases, the BLEU score can have up to 0.5 point difference in the worst case (in most cases there are 0.1~0.3 point difference)."*
- The numbers reported in the paper were indeed rounded figures.

Results on `devtest`

It seems like for the `EN -> *` results, `sacrebleu` needs to be removed from the eval script. Thus, the results are in `BLEU4` format. These correspond to detokenized `sacreBLEU` for `* -> EN` and tokenized BLEU for `EN -> *`, according to the authors (see description of Table 3).

Lang. pair	Reported	Reproduced	Difference
EN-NE	4.3	4.69	0.39
NE-EN	7.6	7.66	0.06
EN-SI	1.2	1.48	0.28
SI-EN	7.2	6.94	0.26

As we can see, these results do differ a bit from what they reported in the paper. However, most of hem seem to fall within a "0.1-0.3" point interval from the reported scores. This is consistent with FAIR's response from above.

Remaining stuff

- I did not reproduce the semi-supervised baseline as those require several GPUs. Given the GPU resources, that could b pursued as well.