Saved Pseudo-Labels

These are the generations of various large models on various large **training** sets. All in all they took about 200 GPU hours to produce.

Available Pseudo-labels

Dataset	Model	Link	Rouge Scores	Notes
XSUM	`facebook/bart-large-xsum`	download	49.8/28.0/42.5	
XSUM	`google/pegasus-xsum`	download	53.3/32.7/46.5	
XSUM	`facebook/bart-large-xsum`	download		Bart pseudolabels filtered to those with Rouge2 > 10.0 w GT.
CNN/DM	`sshleifer/pegasus-cnn-ft-v2`	download	47.316/26.65/44.56	do not worry about the fact that train.source is one line shorter.
CNN/DM	`facebook/bart-large-cnn`	download		5K (2%) are missing, there should be 282173
CNN/DM	`google/pegasus-xsum`	download	21.5/6.76/25	extra labels for xsum distillation Used max_source_length=512, (and all other pegasus-xsum configuration).
EN-RO	`Helsinki-NLP/opus-mt-en-ro`	download		
EN-RO	`facebook/mbart-large-en-ro`	download		

(EN_RO = WMT 2016 English-Romanian).

Example Download Command:

```
curl -S https://cdn-datasets.huggingface.co/pseudo/xsum/bart_xsum_pl.tgz | tar -xvz -C .
```

Generating New Pseudolabels

Here is the command I used to generate the pseudolabels in the second row of the table, after downloading XSUM from here.

```
python -m torch.distributed.launch --nproc_per_node=8 run_distributed_eval.py \
    --model_name google/pegasus-xsum \
    --save_dir pegasus_xsum \
    --data_dir xsum \
    --bs 8 --sync_timeout 60000 \
    --max_source_length 512 \
    --type_path train
```

- These commands takes a while to run. For example, `pegasus_cnn_cnn_pls.tgz` took 8 hours on 8 GPUs.
- Pegasus does not work in fp16 :(, Bart, mBART and Marian do.
- Even if you have 1 GPU, `run_distributed_eval.py` is 10-20% faster than `run_eval.py` because it uses `SortishSampler` to minimize padding computation.

Contributions

Feel free to contribute your own pseudolabels via PR. Add a row to this table with a new google drive link (or other command line downloadable link).