Whole Word Mask Language Model

These scripts leverage the 🤗 Datasets library and the Trainer API. You can easily customize them to your needs if you need extra processing on your datasets.

The following examples, will run on a datasets hosted on our hub or with your own text files for training and validation. We give examples of both below.

The BERT authors released a new version of BERT using Whole Word Masking in May 2019. Instead of masking randomly selected tokens (which may be part of words), they mask randomly selected words (masking all the tokens corresponding to that word). This technique has been refined for Chinese in this paper.

To fine-tune a model using whole word masking, use the following script:

```
python run_mlm_wwm.py \
    --model_name_or_path roberta-base \
    --dataset_name wikitext \
    --dataset_config_name wikitext-2-raw-v1 \
    --do_train \
    --do_eval \
    --output_dir /tmp/test-mlm-wwm
```

For Chinese models, we need to generate a reference files (which requires the ltp library), because it's tokenized at the character level.

Q : Why a reference file?

A : Suppose we have a Chinese sentence like: `我喜欢你` The original Chinese-BERT will tokenize it as `['我','喜','欢','你']` (character level). But `喜欢` is a whole word. For whole word masking proxy, we need a result like `['我','喜','##欢','你']`, so we need a reference file to tell the model which position of the BERT original token should be added `##`.

Q : Why LTP ?

A : Cause the best known Chinese WWM BERT is Chinese-BERT-wwm by HIT. It works well on so many Chines Task like CLUE (Chinese GLUE). They use LTP, so if we want to fine-tune their model, we need LTP.

You could run the following:

```
export TRAIN_FILE=/path/to/train/file
export LTP_RESOURCE=/path/to/ltp/tokenizer
export BERT_RESOURCE=/path/to/bert/tokenizer
export SAVE_PATH=/path/to/data/ref.txt

python run_chinese_ref.py \
    --file_name=$TRAIN_FILE \
    --ltp=$LTP_RESOURCE \
    --bert=$BERT_RESOURCE \
    --save_path=$SAVE_PATH
```

Then you can run the script like this:

```
export TRAIN_FILE=/path/to/train/file
export VALIDATION_FILE=/path/to/validation/file
export TRAIN_REF_FILE=/path/to/train/chinese_ref/file
export VALIDATION_REF_FILE=/path/to/validation/chinese_ref/file
export OUTPUT_DIR=/tmp/test-mlm-wwm

python run_mlm_wwm.py \
    --model_name_or_path roberta-base \
    --train_file $TRAIN_FILE \
    --validation_file $VALIDATION_FILE \
    --train_ref_file $TRAIN_REF_FILE \
    --validation_ref_file $VALIDATION_REF_FILE \
    --do_train \
    --do_eval \
    --output_dir $OUTPUT_DIR
```

Note1: On TPU, you should the flag `--pad_to_max_length` to make sure all your batches have the same length.

Note2: And if you have any questions or something goes wrong when runing this code, don't hesitate to pin @wlhgtc.