

## Motivation

Without processing, english-> romanian mbart-large-en-ro gets BLEU score 26.8 on the WMT data. With post processing, it can score 37.. Here is the postprocessing code, stolen from @mjpost in this [issue](#)

## Instructions

Note: You need to have your test\_generations.txt before you start this process. (1) Setup `mosesdecoder` and `wmt16-scripts`

```
cd $HOME
git clone git@github.com:moses-smt/mosesdecoder.git
cd mosesdecoder
git clone git@github.com:rsennrich/wmt16-scripts.git
```

(2) define a function for post processing. It removes diacritics and does other things I don't understand

```
ro_post_process () {
    sys=$1
    ref=$2
    export MOSES_PATH=$HOME/mosesdecoder
    REPLACE_UNICODE_PUNCT=$MOSES_PATH/scripts/tokenizer/replace-unicode-
punctuation.perl
    NORM_PUNC=$MOSES_PATH/scripts/tokenizer/normalize-punctuation.perl
    REM_NON_PRINT_CHAR=$MOSES_PATH/scripts/tokenizer/remove-non-printing-char.perl
    REMOVE_DIACRITICS=$MOSES_PATH/wmt16-scripts/preprocess/remove-diacritics.py
    NORMALIZE_ROMANIAN=$MOSES_PATH/wmt16-scripts/preprocess/normalise-romanian.py
    TOKENIZER=$MOSES_PATH/scripts/tokenizer/tokenizer.perl

    lang=ro
    for file in $sys $ref; do
        cat $file \
        | $REPLACE_UNICODE_PUNCT \
        | $NORM_PUNC -l $lang \
        | $REM_NON_PRINT_CHAR \
        | $NORMALIZE_ROMANIAN \
        | $REMOVE_DIACRITICS \
        | $TOKENIZER -no-escape -l $lang \
        > $(basename $file).tok
    done
    # compute BLEU
    cat $(basename $sys).tok | sacrebleu -tok none -s none -b $(basename $ref).tok
}
```

(3) Call the function on test\_generations.txt and test.target For example,

```
ro_post_process enro_finetune/test_generations.txt wmt_en_ro/test.target
```

This will split out a new blue score and write a new fine called `test_generations.tok` with post-processed outputs.

