使い方

1. henkan.py
   漢字交じり文を「漢字交じり文」「ひらがな文」「単語別漢字交じり文」「単語別ひらがな文」の 4 つに変換し、csv ファイルにしています。
2. data_set.py
   1 で作成した csv ファイルを train 用、test 用、dev 用に分割します(0.9 : 0.05 : 0.05)
   data フォルダの中にファイルが作られます。
3. input.py
   学習を行います。自身は GoogleColab を使用し計 10 万文を読み込ませました。GPU ありで約 5 時間ほどかかりました。T5 を用いています。

参考

   https://github.com/Bandolu/Hogen/blob/master/GPU_Hougen_T5.ipynb
   https://github.com/sonoisa/t5-japanese

4. output.py
   実際にひらがな文から漢字交じり文を予測生成します。学習で作成した model を読み込み、文を生成できます。

ライブラリ系

torch==2.0.*
torchtext==0.15.*
torchvision==0.15.*
torchaudio==2.0.*
torchmetrics==0.11.*
torchdata==0.6.*
transformers==4.26.1
pytorch_lightning==1.9.3
sentencepiece==0.1.97