

おわりに

古典中国語係り受け解析システム GuwenCOMBO と、日本語係り受け解析システム UniDic-COMBO を、開発・公開した。共通テスト『国語』による評価の結果をまとめると

- 第4問 古典中国語 BERT モデルは、たとえ簡化字で作られていても、古典中国語の係り受け解析に効く
- 第1問 Wikipedia をもとにした現代日本語 BERT モデルは、現代日本語の係り受け解析に効く
- 第2問 Wikipedia をもとにした現代日本語 BERT モデルは、近代日本語の係り受け解析にあまり効かない
- 第3問 Wikipedia をもとにした現代日本語 BERT モデルは、中古和文の係り受け解析に全く効かない。

ということになるだろう。端的に言えば、適材適所ということだ。

ただ、近代日本語や中古和文の BERT モデルを作成するのは、正直なところ、われわれの手に余る。次善策として、UniDic 見出し語をもとにした NWJC-BERT^[52]の採用も考えたが、しかし、NWJC-BERT は有償配布^[53]である。UniDic-COMBO に組み込むわけにはいかないだろう。近代日本語や中古和文の係り受け解析に対しては、しばらくは UniDic2UD で我慢するしかない、ということである。

^[52]浅原正幸, 西内沙恵, 加藤祥: NWJC-BERT: 多義語に対するヒトと文脈化単語埋め込みの類似性判断の対照分析, 言語処理学会第 26 回年次大会発表論文集 (2020 年 3 月), pp.961-964.

^[53]<https://www.gsk.or.jp/catalog/gsk2020-e/>