学際大規模情報基盤共同利用・共同研究拠点公募型共同研究 2022 年度採択課題

jh221007 安岡孝一 (京都大学人文科学研究所附属東アジア人文情報学研究センター)

単語間に区切りのない書写言語における係り受け解析エンジンの開発

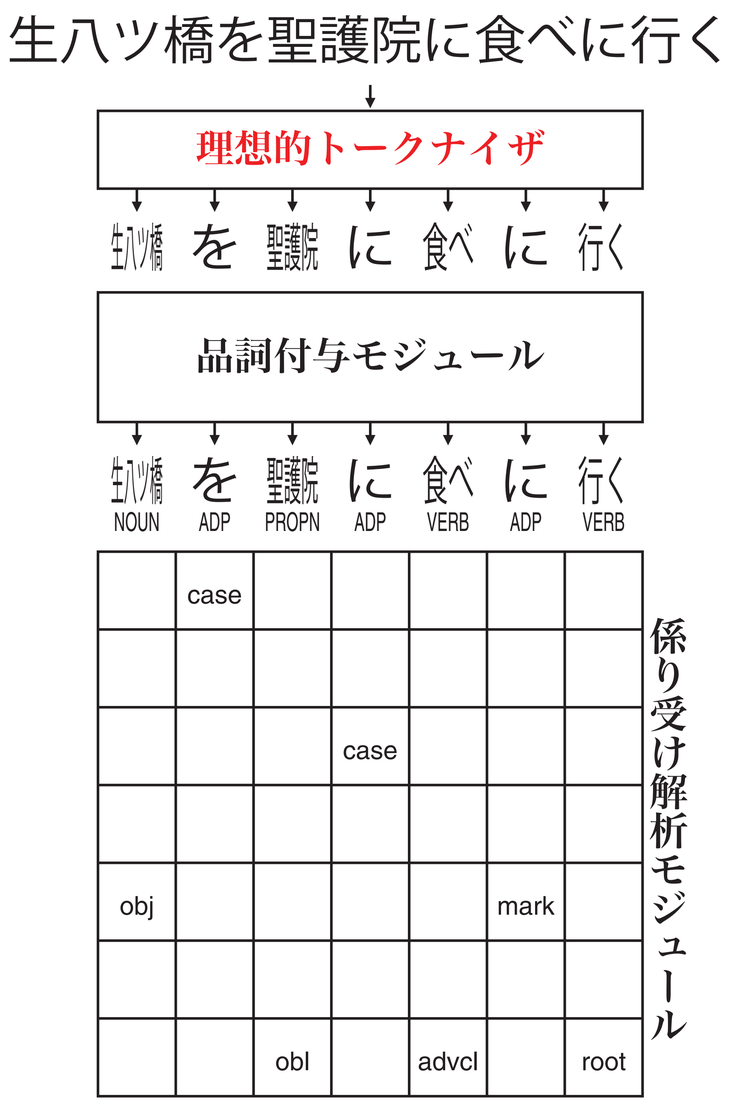
山崎直樹・二階堂善弘 (関西大学), 師茂樹 (花園大学), 鈴木慎吾 (大阪大学), Christian Wittern・池田巧・守岡知彦・白須裕之・藤田一乘 (京都大学)

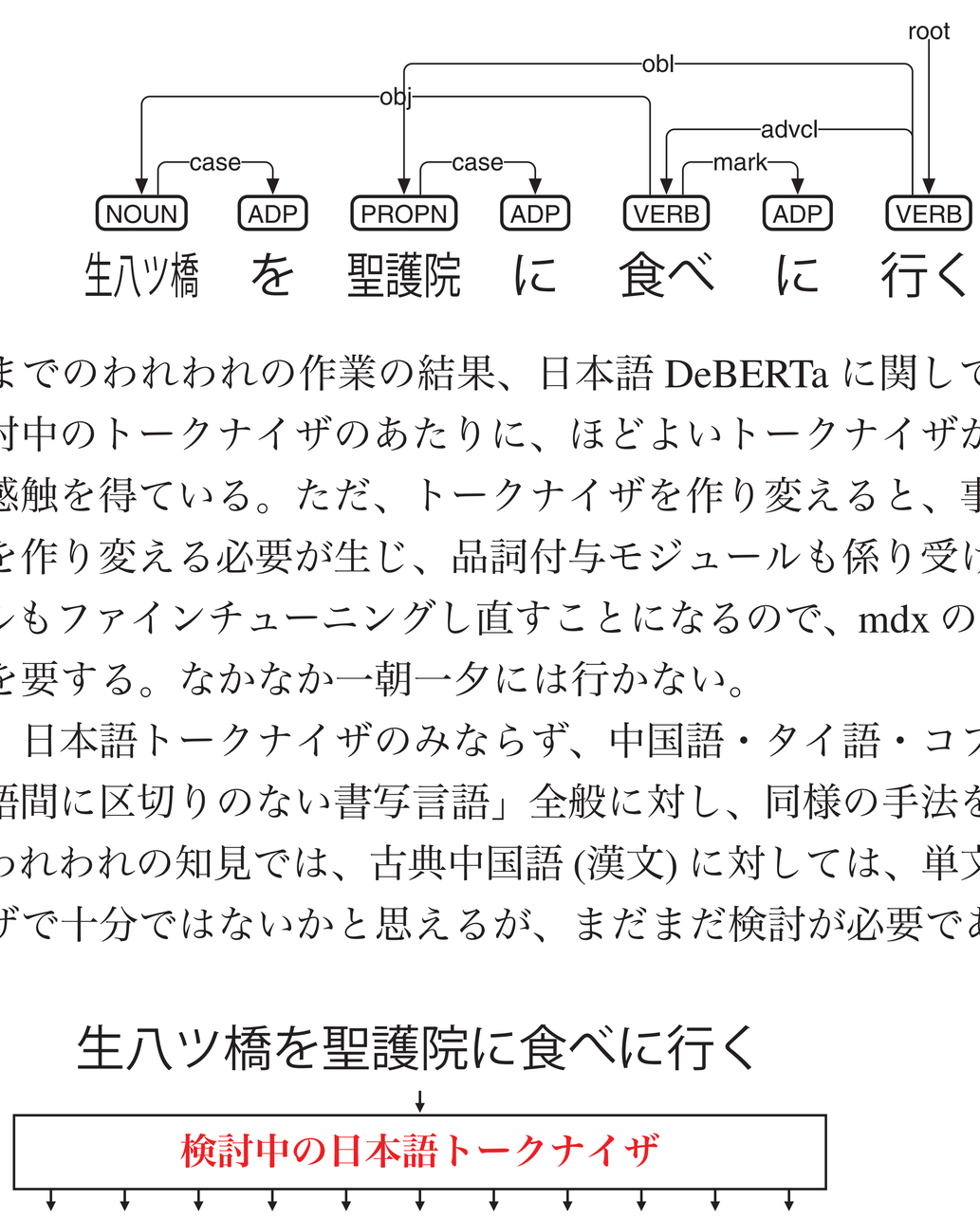
BERT / RoBERTa / DeBERTa 等の事前学習モデルを用いた係り受け解析エンジンは、

1. トークナイザ (tokenizer)
2. 品詞付与モジュール (part-of-speech tagger)
3. 係り受け解析モジュール (dependency parser)

の 3 段構成で設計される。これらのうち、品詞付与モジュールには Conditional Random Fields が、係り受け解析モジュールには Biaffine が、それぞれファインチューニング手法として高い精度を上げている。一方でトークナイザは、かなり言語依存であるらしく、特に日本語・中国語・タイ語など「単語間に区切りのない書写言語」に関しては、解析精度との関係がまだ十分に解明されていない。

日本語 (国語研長単位 Universal Dependencies) を例にとると、単文字トークナイザ (右図) は設計が容易なものの、トークンの種類が少なくなりすぎる上に、各モジュールが肥大化してしまう。単語とトークンを完全に合致させた理想的トークナイザ (左下図) は、トークンの種類が多くなりすぎる上に、設計が非常に困難である。これらは両極端なトークナイザの例であり、トレードオフを考えるなら、これらの間に、トークンの種類数も、モジュールの規模も、ほどよいトークナイザが存在するはずである。そのようなトークナイザを突き止めたい。





これまでのわれわれの作業の結果、日本語 DeBERTa に関しては、下図で検討中のトークナイザのあたりに、ほどよいトークナイザが存在しそうな感触を得ている。ただ、トークナイザを作り変えると、事前学習モデルを作り変える必要が生じ、品詞付与モジュールも係り受け解析モジュールもファインチューニングし直すことになるので、mdx の GPU でも数日を要する。なかなか一朝一夕には行かない。

また、日本語トークナイザのみならず、中国語・タイ語・コプト語など「単語間に区切りのない書写言語」全般に対し、同様の手法を研究したい。われわれの知見では、古典中国語 (漢文) に対しては、単文字トークナイザで十分ではないかと思えるが、まだまだ検討が必要である。