公開日本語言語モデルとその評価の現状

林 政義
株式会社ワークスアプリケーションズ・システムズ

あらまし

- ワークス徳島人工知能NLP研究所から日本語言語モデル chiTra を公開
  - モデルの詳細については本大会発表 (PT1-6) やGithubを見ていただきたい
    - http://github.com/WorksApplications/SudachiTra
- これにあたって事前調査した日本語言語モデルの現状についてのお話
  - どれくらいあるか・それぞれの特徴は・どうやって選べばよいのか
  - 新たに公開するにあたって何を用意しておくべきか
- 簡易まとめ
  - 現状モデルの評価や比較の基準となるものがほぼない
  - chiTraモデルでの方針が参考になればと思います

公開日本語言語モデル

	公開	コーパス	評価タスク（公開データセット）
Google (Multilingual BERT)	2018/11	Wikipedia	
yoheikikuta	2019/01	Wikipedia	
ホットリンク	2019/03	日本語ツイート	
京都大学	2019/04	Wikipedia	
ストックマーク	2019/04	ビジネスニュース記事	
東北大学	2019/11	Wikipedia	
NICT	2020/03	Wikipedia	
Laboro.AI	2020/04	ウェブコーパス	
akirakubo	2020/08	Wikipedia + 青空文庫	
レトリバ	2021/04	日本語話し言葉コーパス	

...

公開日本語言語モデル

	公開	コーパス	評価タスク（公開データセット）
Google (Multilingual BERT)	2018/11	Wikipedia	XNLI
yoheikikuta	2019/01	Wikipedia	livedoor ニュース
ホットリンク	2019/03	日本語ツイート	ツイッター評判分析
京都大学	2019/04	Wikipedia	京大テキストコーパス構文解析
ストックマーク	2019/04	ビジネスニュース記事	-
東北大学	2019/11	Wikipedia	-
NICT	2020/03	Wikipedia	解答可能性付き読解
Laboro.AI	2020/04	ウェブコーパス	livedoor ニュース, 運転ドメインQA
akirakubo	2020/08	Wikipedia + 青空文庫	-
レトリバ	2021/04	日本語話し言葉コーパス	-

...

chiTraモデルの評価

- BERT論文での性能評価
  - GLUE, SQuAD, SWAG
- → タスク種が類似・他で利用されているデータセットを採用
  - Multilingual Amazon Review （日本語部分）
    - 評判分析、JGLUEへの採用予定（昨年時点）
  - 京都大学常識推論
    - 常識推論、JGLUEへの採用予定（昨年時点）
  - 解答可能性付き読解
    - 読解、NICT-BERTの評価実験

まとめ

- 日本語言語モデルは多数公開されているが共通の評価基準はない状態
  - データセットの整備とともにこちらの指標も発展することを期待
- chiTraモデルの評価が参考になれば
  - Multilingual Amazon Review, 京大常識推論, 解答可能性付き読解
  - モデル・スクリプトを公開中
    - http://github.com/WorksApplications/SudachiTra