【別　紙】

＜ Laboro.AI BERT モデルの精度評価 ＞

Laboro.AI BERT モデルの性能を評価するため、今回、以下２つのタスクで検証を行いました。

＜ タスク（A）　文章分類 ＞

NHN Japan 株式会社が収集し、クリエイティブ・コモンズライセンスのもと公開している livedoor ニュースのコーパス※を用い、特定のニュース記事を９つのカテゴリー（トピックニュース、Sports Watch、IT ライフハック、家電チャンネル、MOVIE ENTER、独女通信、エスマックス、livedoor HOMME、Peachy）に正しく分類できるかを検証・評価しました。

※livedoor ニュースコーパスについてはこちらをご覧ください。http://www.rondhuit.com/download.html#ldcc
※livedoor は NHNJapan 株式会社の登録商標です。

＜ タスク（B）　質問回答 ＞

与えられた文章の中から質問に対する答えを抽出・回答するタスクで、正しい回答ができるかの精度を評価しました。今回は「運転ドメイン QA データセット※」という、インターネット上で公開されている運転に関するブログ記事を元に構成されたデータセットのうち、文章読解のための Q&A データセットである「RC-QA データセット」というものを引用しています。例えば、

・文章　：私の車の前をバイクにまたがった警察官が走っていた。
・質問　：警察官は何に乗っていた？
・答え　：バイク

といった一群がセットになっています。

※「運転ドメイン QA データセット」は、京都大学大学院 情報学研究科 黒橋禎夫教授・河原大輔准教授・村脇有吾助教 研究室が公開するものです。詳しくはこちらをご覧ください。http://nlp.ist.i.kyoto-u.ac.jp/index.php?Driving%20domain%20QA%20datasets

＜ 精度評価　＞

上記の２つのタスクそれぞれについて、以下の３つのモデルでその精度を比較しました。

① 公開されている日本語版 Wikipedia のコーパスを事前学習させたモデル※
② Laboro.AI BERT Base モデル　（12 層、ハイパーパラメーター数 110M）
③ Laboro.AI BERT Large モデル（24 層、ハイパーパラメーター数 340M）

複数回の検証結果を平均した比較表がこちらの次表です。

	コーパスサイズ (corpus size)	タスク（A） 文章分類の正解率 (accuracy)	タスク（B） 質問回答の一致率 (exact match)
①　日本語版 Wikipedia モデル	2.9GB	97.2%	76.3%
②　Laboro.AI BERT Base モデル	12GB	97.7%	75.5%
③　Laboro.AI BERT Large モデル	12GB	98.1%	77.3%

タスク（A）文章分類・タスク（B）質問回答ともに、いずれのモデルも僅差で高い精度を示している中、③ Laboro.AI BERT Large モデルがとくに高い結果を示していることが確認できました。

※日本語版 Wikipedia のコーパスを事前学習させたモデルとしては、「BERT with SentencePiece for Japanese text」(Yohei Kikuta 氏、https://github.com/yoheikikuta/bert-japanese)で公開されているものを使用。