

学生・若手研究者のためのBERTワークショップ 参加報告

リーダーボード as of 2022/08/09 16:20 NLPオールドタイマーズさん  

ログイン・参加登録

TASK 1: 分類

TASK 2: 属性値抽出

TASK 3: リンキング

Rank	Team Name	Submitted on	Description	Micro-F1 (Public) ↓
1	Kosuke Takemoto	2022/08/09	2019の教師ありデータ bert > lstm > linear, epoch=2	95.7173
2	運営-中山	2022/06/01	RoBERTaベースライン https://github.com/k141303/Shinra2022	95.2224
3	森羅2022実行委員会	2022/06/24	ベースラインシステム	94.2519
4	Yuanzhi Ke	2022/08/08	bert-large-japanese、LAMB、入力長128、エポック8、バッチ32、lr1e-3	89.3139
5	fumikawa	2022/08/08	bert-large-japanese M_LEN=128 EPOCH=6	89.2308
5	Team久山田	2022/08/05	cl-tohoku/bert-base-japanese-v2 MaxLength=256	89.2308
7	Hiroki Yamauchi	2022/08/04	cl-tohoku/bert-base-japanese-whole-word-masking使用	89.1476
8	NLPオールドタイマーズ	2022/08/08	ベースラインで提供されたNotebookでパラメータ変更	88.8981
8	Akira Ogawa	2022/08/04	ワークショップ 最大文長=256 学習率=5e-5 epoch=8	88.8981
10	PS	2022/08/05	MAX_LENGTH = 512, EPOCH = 16	88.6486

Rows per page: 10 1-10 of 48  

上田良寛

自己紹介

上田良寛 yoshihiroueda@gmail.com

■1982年京都大学大学院工学研究科修士課程修了

■1982年 富士ゼロックス入社～2022年 富士フイルムビジネスイノベーション退職

- 1984～機械翻訳システムの研究：LFGベース機械翻訳、MuSystemへの参画
- 1988～ATR自動翻訳電話研究所に出向、機械翻訳システムの中で英文生成部分を担当
- 1991～富士ゼロックス復帰、自然言語処理技術の研究：推敲支援、検索、情報の分類／可視化、要約、複数文書要約
- 2000～開発部門に異動、検索、要約、情報可視化の商品化
- 2006～自然言語処理分野以外の技術開発、サービス開発、サービス運用
- 2017～定年再雇用と同時に研究部門に移籍、言語データの整備など担当

今回参加したモデルの内容

■ハイパーパラメーター変更のみ

- BERTへの入力長 `MAX_LENGTH = 256` # ← 128
- オプティマイザーの学習率 `LEARNING_RATE = 5e-5` # 元と同じ
- 学習回数 `EPOCH = 8` # ← 4
- 学習時のバッチサイズ `BATCH_SIZE = 32` # ← 16

今後の課題

■分類タスクの高精度化：ハイパーパラメーターの調整以外に何があるか？

- 対象とするテキストの範囲の限定: 概要、Infobox、他にエンティティ種類別に向けた内容があるか
- ハイパーリンクの利用
- BERTで潜在的に理解しているとみえるもの: 構文、意味関係、単語の重要度、エンティティ
 - [Jawahar+ 2019] What does BERT learn about the structure of language? <https://aclanthology.org/P19-1356.pdf>

■森羅の上位目的に合わせた処理

- RDFによる構造化: 関係の体系化
- 上位下位概念、集合演算、否定、仮定/反実仮想を含めた推論