学術ドメインに特化した日本語マスク言語モデルの試作

山内 洋輝　梶原 智之（愛媛大学）桂井 麻里衣（同志社大学）大向 一輝（東京大学／国立情報学研究所）二宮 崇（愛媛大学）

1．概要

- 日本語の論文抄録からAcademic RoBERTaを学習
- 2013年から2017年までの科研費の採択課題から研究課題名に関する評価実験
- Academic RoBERTaは汎用的な既存のマスク言語モデルを上回る性能

2．コーパス作成

学術データベースCiNii Articlesの論文抄録を抽出（126万文書）

1. 定型表現の削除（114万文書）　ノイズを含んだ論文抄録の削除
2. 文分割（730万文）　ルールベースの文分割
3. 日本語文の抽出（668万文）　文字単位で半分以上が日本語の文を抽出
4. 重複文の削除（633万文）　重複した文を削除し、コーパスにその文が1回だけ含む
5. 文字数制限（627万文）　極端な短文および長文の削除

3．語彙について

Academic RoBERTaの語彙
- 単語分割を行わずに直接SentencePieceで32,000の語彙を作成

既存のマスク言語モデルの語彙との違い
- 論文表現や専門用語が含まれる
- 語彙の49.4%が既存モデルには含まれていない

論文表現	専門用語
する手法を提案する	ニューラルネットワーク
であることが確認された	ヘモグロビン
について考察を行った	肝障害

4．評価実験：科研費の研究課題名の分類

著者同定（訓練用：100,000文対／検証用：10,000文対／評価用：10,000文対）
- 二つの研究課題が同一著者か否かを分類

文書分類（訓練用：70,000文／検証用：1,500文／評価用：1,500文）
- 研究課題名から4段階の階層構造を持つ研究分野を分類

	著者同定 (acc)	文書分類 (acc)			
クラス数	2	4	14	77	318
東北大BERT	95.1	83.7	69.6	53.3	40.3
早大RoBERTa	97.1	83.9	71.9	55.4	42.7
Academic RoBERTa	**98.7**	**84.7**	**72.9**	**58.8**	**44.6**

- Academic RoBERTaはすべてのタスクにおいて最高性能を達成
- 特に、詳細な専門知識が必要な小区分で大きな性能の改善