Q11-4　学術ドメインに特化した日本語事前訓練モデルの構築

山内 洋輝　梶原 智之（愛媛大）桂井 麻里衣（同志社大）大向 一輝（東大 / NII）二宮 崇（愛媛大）

1．概要

- 日本語の論文抄録からAcademic RoBERTaとAcademic BART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とAcademic BART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　ノイズを含む論文抄録の削除（114万文書）
2. 文分割　ルールベースの文分割（730万文）
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　極端な短文および長文の削除（627万文）

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

著者同定
二つの研究課題が同一著者か否かを分類

文書分類
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

ヘッドライン生成
研究概要からタイトルを生成する要約タスク

	層数	次元数	著者同定 (acc) 2クラス	文書分類 (acc) 4クラス	14クラス	77クラス	318クラス
東北大BERT	6	768	95.1	83.7	69.6	53.3	40.3
早大RoBERTa	6	768	97.1	83.9	71.9	55.4	42.7
Academic RoBERTa	6	768	**98.7**	**84.7**	**72.9**	**58.8**	**44.6**

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

	層数	次元数	ROUGE-1	ROUGE-2	ROUGE-L
京大BART (Base)	6	768	26.0	13.1	23.6
京大BART (Large)	12	1,024	36.7	18.5	33.0
Academic BART	6	768	**43.0**	**21.7**	**38.8**

生成したタイトルを二つの指標で人手評価

	妥当性	流暢性
参照文	4.29	**4.95**
京大BART (Large)	3.63	3.77
Academic BART	**4.59**	4.89

- AcademicBARTはモデルサイズの大きい京大BART（Large）以上の性能を発揮
- 人手評価では参照文に対して提案モデルが流暢性で匹敵しており、妥当性では上回る

研究概要

近年の機械翻訳の進歩に大きく貢献する技術として、対訳データから翻訳ルールを自動的に学習する統計的機械翻訳（SMT）がある。しかしSMTは明示的にある翻訳結果が選ばれる訳の根拠を考慮していないため、人手で構築するルールベース機械翻訳(RBMT)に比べて一部の入力文や言い回しに対して致命的な誤訳を生成することがある。本研究は、人間の翻訳者が用いる根拠を持つ訳選択ルールを自動発見し、SMTと融合する方法の開発を行った。具体的には、言語的情報に基づく翻訳システムの開発、訳選択の根拠の導入、SMTシステムの分析を効率化する枠組みの開発、多言語データにおける訳選択の根拠を発見する手法の開発を行った。

参照文

訳選択の根拠の自動推定とその機械翻訳における応用

Academic BARTの生成文

統計的機械翻訳における訳選択ルールの自動発見

京大BART (Large) の生成文

言語的情報に基づく統計的機械翻訳システムの開発

https://github.com/EhimeNLP/AcademicRoBERTa　https://github.com/EhimeNLP/AcademicBART