P8-14

エンティティの言語間リンクに基づく多言語モデルの構築

西川 荘介　山田 育矢　鶴岡 慶雅　越前 功

研究概要

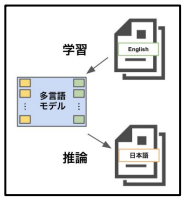
言語間転移タスク において知識ベースに紐づく**エンティティ表現を付加的な特徴量**として用いることで **汎用多言語モデルの性能向上** を実現

研究背景

言語間転移タスク

資源の豊富な言語（原言語）の教師データで学習したモデルを他の言語（目的言語）に適用するタスク

→ 近年は汎用多言語モデルを利用した研究が盛んに



既存手法の問題点①

汎用多言語モデルは以下の場合性能が十分に発揮されない

- 言語対の言語体系が異なる場合
- 事前学習データに含まれない言語への転移

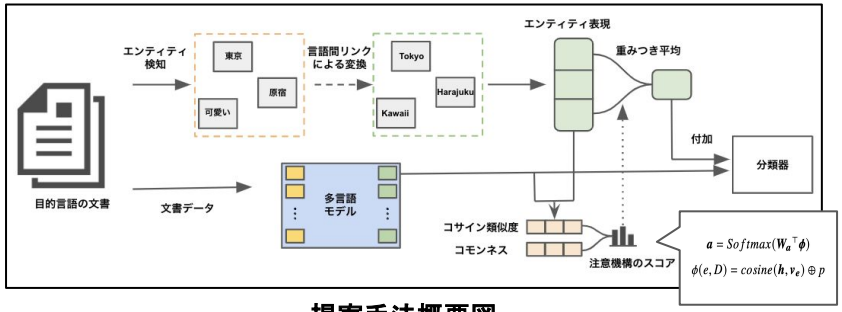
既存手法の問題点②

目的言語のラベルなしデータで追加学習する手法が提案されているが以下の問題点がある

- 目的言語ごとに追加の学習データを用意しモデルを再学習する必要がある
- 複数言語で性能が発揮できるモデルの多言語性が失われる

提案手法

文書中にある知識ベース(Wikipedia)に紐づくエンティティ群を言語間転移タスクにおける付加的な特徴量として利用



提案手法概要図

エンティティの処理

- エンティティ名（e.g."東大"）に紐づくエンティティ群（e.g."東京大学"）を全て抽出
- 推論時のみ言語間リンクを用いて目的言語のエンティティを原言語に変換
- エンティティに対して知識ベースから事前に学習した表現を割り当て

注意機構の導入

- 文書に関与するエンティティを優先する注意機構を同時に学習
- 注意機構の重みは多言語モデルから得られる文表現とエンティティ表現のコサイン類似度・コモンネスから学習

※コモンネス: エンティティ名が特定のエンティティを示す確率

提案手法の利点

- エンティティは曖昧性がなく文書に関連する情報を持つ
- 言語間リンクにより原言語のエンティティ表現を直接推論に利用可能（言語対によらない）
- 目的言語ごとに追加の文書データの収集の必要がなく,単一モデルのみで性能向上が可能（多言語性を失わない）
- 学習済みの多言語モデルに用易に適用可能

実験

実験設定

- Multilingual-BERT [Devlin+,2019]をベースラインとして実験
- 文書分類（MLDoc [Schwenk+,2018]），エンティティ型推定タスク（SHINRA2020-ML [Sekine+,2020]）で評価
- 原言語を英語 ,目的言語をそれぞれ 7言語,10言語として検証
  エンティティ表現は Wikipedia2Vec [Yamada+, 2020]で事前に学習

実験結果

- 多くの言語対でベースラインと比較し性能向上
- 事前学習したエンティティ表現と注意機構が効いている

文書分類の結果（正解率）

モデル	fr	de	ja	zh	it	ru	es	ave.
M-BERT	79.8	74.6	70.2	72.7	67.2	66.4	75.7	75.1
M-BERT + BoE (-embedding)	80.9	76.1	71.3	**76.5**	68.0	67.5	76.2	76.3
M-BERT + BoE (-attention)	77.4	72.2	68.4	71.8	66.0	64.6	74.6	73.6
M-BERT + BoE (full)	**83.6**	**79.2**	**71.5**	74.4	**71.0**	**68.4**	**77.4**	**77.5**

エンティティ型推定の結果（F1値のマイクロ平均）

モデル	ave.
M-BERT	75.7
M-BERT + BoE (-embedding)	78.1
M-BERT + BoE (-attention)	54.2
M-BERT + BoE (full)	**78.6**