進捗報告

1 今週あったこと

ipsj 関西支部大会

2 hottoSNS-BERT に関するまとめ

2.1 概要

- Sakaki [1] らによる大規模日本語 SNS コーパスによる文分散表現モデル
- 収集文の多様性が大きくなるように工夫. (bot 投稿, リツイートの除外, 重複ツイート文の除外)
- 構築されたコーパスのツイート数は 8500 万 (本家 英 Wiki の 35 % 程度の大きさ)
- 特殊トークンとして $< mention >$, $< url >$
- 分かち書きには sentencepiece を採用
- 事前学習として next sentence prediction はツイートに適用することが難しいため, masked language model のみを適用している
- 事前学習のタスク設定について，各サンプルの token 数を最大 64 に制限 (本家は 512)

3 評価

日本語ツイートの評判分析 (Positive/Negative/Neutral の 3 値分類) の結果, 表より, 日本語 Wikipedia のモデルを上回った. 詳細はリンク [1]を参照のこと.

[1] https://github.com/hottolink/hottoSNS-bert

	Twitter 大規模カテゴリコーパス		Twitter 日本語評判分析データセット	
モデル名	accuracy	F-value	accuracy	F-value
BERT MultiLingual	0.7019	0.7011	0.8776	0.7225
BERT 日本語 Wikipedia	0.7237	0.7239	0.8790	0.7359
hottoSNS-BERT	0.7387	0.7396	0.8880	0.7503

4 使い道

漫画は比較的短文で形成されていることや口語表現が多いことから, より合理的なベクトル化が期待できる. 特殊トークンの $< mention >$ も会話の流れを理解させるのに有効であると推測できる.

Tensowflow のモデルで与えられるらしいが, 最近は huggingface-transformer ライブラリの中で Pytorch 用に変換できるらしい. 要確認.

5 今後の予定

- 今後の研究計画を練る
- セリフからの話者推定
- コマ画像のセリフ部分の白抜き

参考文献

[1] Sakaki, Takeshi, Sakae Mizuki, and Naoyuki Gunji. Bert pre-trained model trained on large-scale japanese social media corpus. 2019.