進捗報告

1 研究テーマ・やりたいこと

今のところは, 4 コマ漫画の感情推定を続けてみたい. そのためにやらなければいけないことのピックアップ

- Self-Attention や Transformer を理解してから BERT を動かす環境を構築する
- データセットの再整備 ・ 追加 (2 タッチ 30 話まであるらしい)
- マルチモーダルな解析の考案 (画像処理)

2 読んだ論文

日本語文章のための話し言葉・書き言葉変換コーパス [1]

3 要約

話し言葉を高い可読性を持つ書き言葉へ変換するためには, まず, 話し言葉と書き言葉の文対を大量に用意する必要があり, そして様々な要素を考慮する必要がある. 例えば，フィラーや言い淀み，冗長表現の削除や句読点の付与が必要であるが, 従来の研究ではこれらを独立に捉え, それぞれの要素に適したコーパスを使用していた. さらに日本語には話し言葉, 書き言葉それぞれに特有の表現 (助詞の省略, 文体統一など) が存在するためこれらの問題を無視することはできない.

本稿では, 話し言葉テキストの可読性を向上させることを目的とし, 従来の要素と日本語特有の要素の両方を考慮した, 日本語文章のための話し言葉から書き言葉へ変換するコーパスを作成した.

コーパスの作成にあたっては, いくつかの変換ルールを構築し, クラウドソーシングを用いて 4 つのドメインにおける話し言葉, 書き言葉の文対を人によって作成した. そして, 作成したコーパスで学習したニューラル系列変換モデル (1. attention-based encoder decoder network, 2. pointer-generator network) を用いて話し言葉・書き言葉変換における BLEU, ROUGE-L, METEOR 値のベースラインを確認したところ, 高い指標を得られており (特に pointer-generator network), 話し言葉・書き言葉変換のルールについても網羅的に変換できていることが示された.

4 特徴

既存の話し言葉・書き言葉変換コーパスとの大きな違いとして変換に必要な従来の要素と日本語特有の要素をすべて同時に考慮している点が挙げられる.

5 感想

漫画固有の口語表現の言いかえにも流用できる可能性があると思った.

6 次に読むべき論文

- Self-Attention について書かれているもの.
- pointer-generator network について書かれているもの.

参考文献

[1] 増村 亮 庵 愛, 高島 瑛彦. 日本語文章のための話し言葉・書き言葉変換コーパス. **言語処理学会 第 *26* 回年次大会 発表論文集**, 2020.