BERT による分散表現を用いた.....

1 はじめに

近年, 機械学習の発展に伴い, 自然言語処理の分野においても機械学習を用いた手法が大きな成果を上げている. そして, 自然言語処理のタスクの 1 つである文章生成においても, 次々に新しいモデルが公開されており, そのどれもが短文においては人間と遜色ない文の生成をすることが可能となっている. しかし, これらのモデルは単語間の関係性を考慮しているものが多く, これらのモデルで複数文からなる文章を作成したときには, 文同士の関係を考慮しない文章が生成されてしまうため, どこか違和感のある文章になってしまうことが多々ある.

以上を背景として, 本研究では複数文からなる文章の生成という観点から, 文同士の関係性を推定することを目標とした. 文同士の関係性に明確な定義はないが, 文同士の間にある接続詞はそれらを考慮する重要な指標になる. そこで本稿では, 文章間で接続詞がついている文章を扱い, それらの文章の関係を接続詞を基にいくつかの種類に分類して, それを推定して精度を確認した. 文章の分散表現を得る手法としては, 高い精度が示されている汎用言語モデル BERT を用いた.

2 要素技術

2.1 Doc2Vec

Doc2Vec[**?**] は 単語の分散表現を得る手法である Word2Vec[**?**] から発展した, 単語および文の分散表現獲得手法である. Doc2Vec では文書を分散表現に変換するために Word2Vec に Paragraph ID を導入する. Paragraph ID は各文書と紐づいており, 単語の学習時に共に学習される. Doc2Vec では, この Paragraph ID を文の分散表現として見なす. Word2Vec の CBOW を拡張したモデルを Distributed Memory (DM) といい, Skip-gram を拡張したモデルを Distributed Bag-of-Words (DBOW) という.

2.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) [1] は, 複数の双方向 Transformer に基づく汎用言語モデルであり, 2018 年に Google が発表した言語モデルである. これまでの言語モデルは特定の学習タスクに対して 1 つのモデルを用いてきたが, BERT は大規模コーパスに対して事前学習を施して, 各タスクに対してファインチューニングをすることで, さまざまなタスクに柔軟に対応することができる. 事前学習には入力の一部の単語を “[MASK]” に置き換えてその元単語を予測するように訓練するタスクと 2 文を入力としてその連続性を識別するように訓練するタスクが用いられる. 本稿では, 京都大学から公開されている, 日本語 Wikipedia より全 1,800 万文を用いて事前学習されたモデル[1] を使用した. BERT に文章を入力する際には, 文章の先頭の先頭に “[CLS]” トークンを付与する. BERT は単語ごとの分散表現を出力するが, “[CLS]” トークンに対する出力を文章全体の分散表現として扱うことができる. また, 2 文を扱う際には, 文章の間に “[SEP]” トークンを付与する.

3 データセット

3.1 使用データ

本稿では叙述的な文章として毎日新聞データセット[2] の新聞記事を用いた. このデータセットにはジャンルごとに 2008 年から 2012 年までの記事がある. そのなかのジャンルが 123 面のものを用いた.

4 数値実験

4.1 実験準備

4.2 実験 1

4.3 実験 1 結果と考察

5 まとめと今後の課題

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of

[1] http://nlp.ist.kyoto-u.ac.jp/index.php

[2] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.