令和 6 年度 特別研究報告書要旨

日本語文法誤り訂正モデルの提案について

Y210165 中西 健太朗

指導教員 三好 力

1. はじめに

新型コロナウィルスの感染拡大に伴い、オンライン授業やリモートワークが一般的に行われるようになり、zoom や Microsoft teams などのビデオ通話アプリが使われる機会が多くなった。議事録や講義ノートの作成にもアプリ内の自動音声文字起こし機能が使われるようになったが、通話環境などの要因により十分な精度で文字起こしされない場合もある。自動文字起こし機能は話者の録音環境や滑舌などの問題に音声認識の精度が左右されてしまう。そのため、本研究では文字起こしされた文章に含まれる文法的な誤りを訂正することが出来る機械学習モデルを構築することを目的とする。

2. 実験

構築手法は、日本語のテキストデータを使って事前学習されたモデルを追加データでファインチューニングし、文法誤り訂正タスクに適応させるというものである。事前学習済みモデルには東北大学自然言語処理研究グループが公開している BERT モデルを使用し、追加データは京都大学のメディア研究室が公開している日本語 Wikipedia 入力誤りデータセット（v1）を用いる。ファインチューニングには Pytorch Lightning を使用し、チューニング前のモデルとチューニング後のモデルでどれだけ性能が上がっているのか確認する。

3. 実験結果

チューニングの前後でモデルの性能がどれほど変化したのかを以下の表に記す。

表 1 チューニングの結果

エポック数	検出精度	訂正精度
0	24%	11%
5	84%	78%
10	84%	77%

表のとおり、トレーニングによって事前学習済みモデルに比べるとモデルの性能は検出・訂正ともに向上したが、エポック数が 5 回の時と 10 回の時で性能がそれほど変化せず、訂正精度に至っては 1%ではあるが性能が下がってしまった。

4. 終わりに

実験の結果、誤変換された漢字の検出・訂正精度はともに 80%前後に達し、提案した手法にある程度の効果があることが示された。一方で、多くの課題も明らかになり、提案した手法にはさらなる改善が必要であることが分かった。

本研究の結果から、事前学習済み BERT モデルは、日本語文法誤り訂正タスクにおいてある程度の効果を発揮できることが分かった。特に、誤変換の検出や訂正において、事前学習済みモデルの適用が有効であることが確認できた。また、学習率スケジューラやデータローダに入力するトークンの長さなどの調整がモデルの性能に影響を与えることも分かった。