評価データの言い換えに伴う汎用言語モデルの頑健性の検証

学生証番号 644943　　インテリジェントシステム学科　　西浦 大介 （宮森研究室）

1 はじめに

近年、人間と同等レベルの読解力の実現を目指す試みとして機械読解の研究が盛んに行われている。特に、自然言語を扱う汎用的な言語モデルとして2018年に提案されたBERTは、機械読解以外のタスクでも高い性能を示し、そこから派生したモデルが次々と提案されている。その中でもRoBERTaは、BERTの学習方法を改善することでより高い性能を示すことが報告されたモデルであるが、評価に使われたデータセットが数種類であり、より一般的なデータに対してどの程度報告された性能を維持できるのかについては必ずしも明らかではない。そこで、本稿では、自然言語理解タスクに用いられるデータセットの一部を意図的に変更したデータセットを用いて、RoBERTa がどの程度報告された性能を維持する頑健性をもつのか検証する。

2 検証手順

汎用言語モデルRoBERTaに対して、対象となる評価データセットを意図的に変更することで、モデルの頑健性を検証する。まず、評価データセットを、RoBERTaが正答したデータ群と誤答したデータ群に分割する。次に、各々のデータ群の傾向から代表して選んだデータを、意図的に改変したデータ群を作成し、再度 RoBERTa が正答するかどうかを調べる。

データセットとしては、一般常識に基づいた多肢選択質問応答のデータセットであるCommonsenseQA（以下、CSQAと呼ぶ）と、複数の自然言語処理タスクから構成される言語理解評価用データセットであるGLUEを用いた。

意図的な改変は次の手順で行った。CSQAについては、質問のみの変更として、質問の意味を大きく変えないように単語やフレーズを変更・追加し、選択肢のみの変更として、正答以外の選択肢について、より回答が難しくなるように単語を変更した。GLUEについては、まず、文ペアのいずれか1文を選択し、文中の名詞1単語を同義語に変換した(以後、Easyと呼ぶ)。次に、Easyで作成した内容を元に、Easyで変更した箇所以外の品詞を少なくとも1つ以上できるだけ多く同義語に変更した(以後、Hardと呼ぶ)。以上の変更は、正答、誤答データのいずれに対しても同様に行った。

3 評価および考察

CSQAについては、質問のみの変更として、フレーズ単位の変更が加えられた場合に1〜2割程度正解率の変化が見られ(正答は誤答に、誤答は正答になった)、単語単位の変更では変化がみられなかった(正答は正答のまま、誤答は誤答のまま)。また、選択肢のみの変更について、1〜4 割程度正解率が変化することが確認された。

GLUEについては、Easyの変更が加えられた場合、ほとんど正解率の変化が見られなかった。一方、Hardの変更が加えられた場合、2〜5 割程度正解率が変化することが確認された。

評価結果から、RoBERTaは、1 単語の同義語変更だと結果はほとんど変化せず、強い頑健性をもつ一方、複数単語の同義語変更、あるいは、1 つ以上のフレーズの変更だと2〜5 割程度結果が変化すること、特に、多肢選択の場合、選択肢がより紛らわしい内容に変更されると1〜4 割程度結果が変化することがわかった。フレーズ単位の言い換えや、複数単語の変更に伴う意味内容の変化をより的確に捉えるための改善が、今後必要であると考えられる。

4 まとめ

汎用言語モデル RoBERTa について、主に言い換えに伴う頑健性をいくつかのタスクで検証した。