6M-06

大規模言語モデルによる会話の連続性判定手法の提案

林 想汰 †　　小林 亜樹 †
† 工学院大学情報学部情報通信工学科

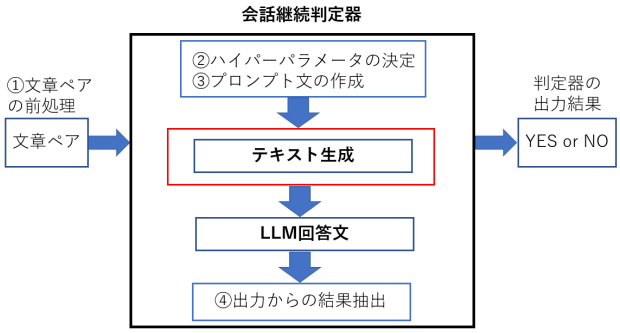

図 1: 提案手法概要図

1　はじめに

SNS などで会話的に交わされるテキストから特定の話題に関する情報を抽出する需要がある．しかし，一般に会話中では所々で話題が転換するため，特定の主題についてのテキストは会話全体の一部である．この会話テキストにおいて話題の転換点を見つけ出せれば主題部分の切り出しに寄与できる．人と対話システムによるやり取りによるデータセット [1] を使った対話破綻の研究はあるが，人同士のやりとりを使ったものはない．そこで本稿では LLM を用いて，SNS に投稿された会話文の主題が同一なのか転換されたかを判定する手法を提案する．

2　提案手法

2.1　概要

LLM に判定させたい会話文を用意し，プロンプト文や出力結果を工夫することで会話において話題が継続しているかを判定する判定器を作成した．概要図として図 1 を示す．

2.2　会話文の用意と前処理

LLM に判定させたい文章ペアを用意する．その際に前処理として文章ペア内にある改行は全て削除する．これは本研究では改行による話題の転換への影響はないものとみなしたためである．

2.3　生成パラメータの決定

LLM はテキストを生成する際に生成する文章に変化を与えるものとしてプロンプト文と生成パラメータがあげられる．そのなかでも生成パラメータを具体的な数値で設定することにより生成させる文章を細かく変えることが可能となる．そこで本研究では生成パラメータの値を 2 つ設定する．1 つ目は生成するテキストのトークン数を制限する max_new_tokens の値を 10 に設定することである．これは生成するテキストのトークン数が大きい場合，判定に必要な情報以外も生成されてしまい判定結果に影響を及ぼしてしまうことを防ぐためである．2 つ目は生成するテキストのランダム性を操作できる temperature の値を 0.1 に設定することである．これは文章ペアを判定させたときの判定結果は現実的な回答を望んでいるという狙いがある．

2.4　プロンプト文の作成

生成されるテキストはプロンプト文の内容によっても左右されるため，いくつかあるプロンプト記法においてどの記法を選びどのように与えるかが重要となる．次に，LLM への指示となるプロンプト文には例示（例題と回答例）を含める shot 文技法が知られている [2]．また，このとき LLM にプロンプトとして与える際に提示する例題のことを判定例と呼ぶ．本論文では異なる shot 文による複数のプロンプト文を作成し，shot 文による回答精度など判定に及ぼす影響について調査する．さらに，判定結果を集計しやすくするために判定結果は YES か NO のどちらかだけが出力されるように指示する．

2.5　出力からの結果抽出

LLM 回答文を処理する際の概要図として図 2 を示す．LLM 回答文は入力文である命令文や判定対象の会話文などを含むプロンプト文の後に続く形で生成される．そのため，判定結果として必要のないプロンプト文，改行コード，削除しても問題の無いモデル特有の生成文を LLM 回答文から削除を行う．図 2 の削除は LLM 回答文内に一致する削除対象のテキストが存在した場合に LLM 回答文から削除する．削除後のテキストを判定結果と呼ぶ．

3　評価実験

3.1　実験目的

本研究で提案した LLM に設定したパラメータの値やプロンプト記法によって会話の継続性判定への有効性を明らかにすることを目的とする．

3.2　実験手順

提案方式による話題継続判定の性能評価を行った．以降では正解ラベルが継続のデータは正例，非継続のデータは負例と呼ぶ．Tweet とその reply である tweet ペア 500 件 (正例 486 件，負例 14 件) について話題が継続しているかを筆者 1 人が人手により正解ラベル付与を行いデータセットとする．使用した LLM のモデルは，Xwin-LM-13B-V0.1[3] である．

Proposal of conversation continuity determination method with large language models
†Souta Hayashi　†Aki Kobayashi
†Department of Information and Communications Engineering, Faculty of Engineering, Kogakuin University

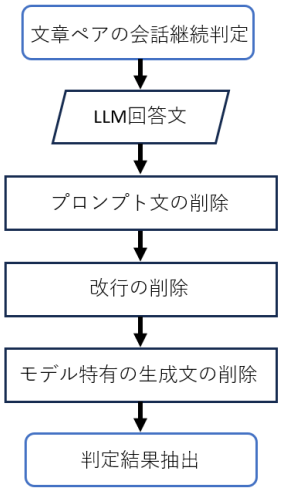

図 2: 出力テキスト処理フローチャート

次に LLM に判定させる際にパラメータの値は提案手法の値を設定した．設定 1 としてプロンプト文に判定結果が継続の例題を 5 つ，設定 2 として判定結果が非継続の例題を 5 つ合計 10 パターンで生成を行った．この時の例題は取得した文書ペアの内，データセットに含めていないものを用いた．

LLM 回答文から判定結果のみを抽出し，正解ラベルとの比較を部分一致にて行った．その後，比較結果を集計する．正例数と負例数に偏りがあるため正例，負例それぞれで判定精度を算出した．精度は正解ラベルとの一致件数を正解ラベルの件数で割って算出する．また比較手法として temperature の値のみを 0.5 に変更して提案手法と同様の手順で設定 3 としてプロンプト文に判定結果が継続の例題を 5 つ，設定 4 として判定結果が非継続の例題を 5 つで実験を行い精度を算出した．

4 結果

図 3 は正例に対して判定を行った際の判定器の精度の結果である．また，図 4 は負例に対して判定を行った際の判定器の精度の結果である．結果のグラフは横軸が例題数，縦軸が判定器における判定精度である．図のデータはの例題数毎に設定 1，設定 2，設定 3，設定 4 の順で並んでいる．図 3 を見ると設定 1 では例題数とともに判定精度も向上し，例題数を 5 個与えたプロンプト時では最も精度が高いことが分かった．次に図 4 を見ると設定 3 では例題数が 3,4,5 と増えると比較手法で同じ判定例を提示した時よりも判定精度も向上することが分かった．

5 考察

設定 1 で例題数とともに判定精度も向上した要因として，1 つの例題ではその例題に近い内容の文章ペアでしか話題が継続しているかの判定をすることはできないが，例題数が増えることで LLM に与える情報量が増えたため精度が向上したのではないかと考えられる．

また，設定 3 では例題数が 3,4,5 と増えると設定 4 よりも判定精度も向上した．この要因としては比較手法

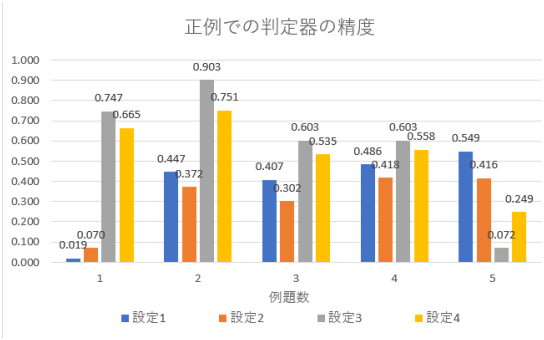

図 3: 正例での判定器の判定精度

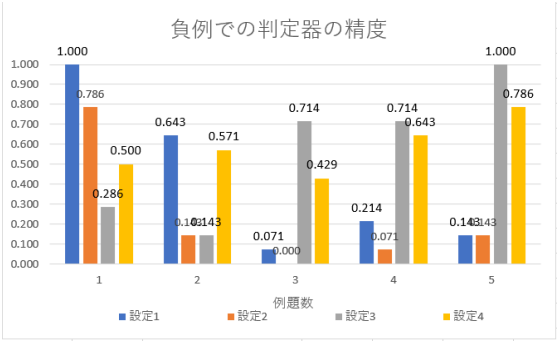

図 4: 負例での判定器の判定精度

では YES も NO もどちらも出力されないや，両方が出力されてしまう判定不能扱いになっている件数が提案手法と比べて多かった．これは比較手法では提案手法の値より temperature の値を高くしたためテキスト生成における自由度が高まり，事実に基づいた回答がしにくくなっているのが要因だと考えられる．

6 おわりに

本研究では LLM を用いた会話継続判定器を提案し，生成パラメータの値やプロンプトの記法を変えて会話継続判定器の評価を行った．生成パラメータ数を変化させた比較手法と比べて提案手法の有効性を示すことができた．しかしデータセット内の正解ラベルの割合に偏りがあり，非継続ラベルを対象とした判定器の精度はデータ数が少ないことから正確な精度とは言えない．そのため非継続ラベルのデータ数を増やすことが今後の課題である．

参考文献

[1] 東中 竜一郎，船越 孝太郎, 小林 優佳, 稲葉 通将，“対話破綻検出チャレンジ，”第 75 回 言語・音声理解と対話処理研究会 (第 6 回対話システムシンポジウム)，人工知能学会研究資料 SIG-SLUD-75-B502 巻，pp.27-32，2016.

[2] Few-Shot Prompting https://www.promptingguide.ai/techniques/fewshot (参照 2024-1-9)

[3] Xwin-LM-13B-V0.1 https://huggingface.co/Xwin-LM/Xwin-LM-13B-V0.1 (参照 2024-1-9)