5V-03

ユーザベクトルの単語選好重み付けによる返答生成モデルのスタイル反映性向上

大石 樹† 渥美 雅保†

創価大学大学院理工学研究科情報システム工学専攻†

1. はじめに

本研究では,事前学習言語モデルの入力に返答者固有の ID と返答生成制御用の特殊トークンを追加しファインチューニングを行うことで,スタイル制御可能な返答生成モデルを構築する. ここで,本研究では「スタイル」という言葉を「個性に由来する単語選択規則」として扱う.そして,ユーザの個性的言語スタイル特徴を反映させるために,ユーザ ID ベクトルに対し単語の使用頻度に基づくユーザの単語選好情報を反映する重み付けを行ない,スタイル反映性の向上を図る.また,返答生成モデルの評価にはテキスト分類による返答者推定を利用したスタイルの客観的評価を行なう.

2. 返答生成モデル

2.1. 事前学習モデル

返答生成モデルは,「GPT-2 日本語モデル[1]」を学習用データセット(3章2節参照)によりファインチューニングして構築する.

2.2. データ形式

返答生成モデルの学習時に使用するデータ形式は

発話<|beforeid|>ユーザ ID 返答<|endoftext|> (1)

である.また,推論時の入力データ形式を

発話<|beforeid|>ユーザ ID (2)

とする. ここで,<|beforeid|>と<|endoftext|>は,それぞれ,発話の終わりと返答の終わりを示す特殊トークンである.

2.3. スタイル制御

2.3.1. スタイル情報の入力方法

スタイル情報を語彙として入力するために,事前学習言語モデルに対し各ユーザ ID 並びに特殊トークン<|beforeid|>の語彙への追加を行う.ユーザ ID ベクトル及び<|beforeid|>用ベクトルは事前学習言語モデルの語彙外のトークンであるため,事前に生成したベクトルを使用する.ユーザ ID ベクトルは,同一ユーザの返答データを文ベクトルに変換し,それらの平均ベクトルをユーザ ID ベクトルと定義する.これにより,ユーザ毎に異なるベクトルを割り当てる.<|beforeid|>ベクトルには値が全て 1 のベクトルを割り当てる.

2.3.2. ユーザ ID ベクトルの重み付け

スタイル反映性向上を目的として,ユーザ ID ベクトルに重み付けをする.重みには,TF-IDF と,TF-IDF を元にユーザ間のスタイルの違いの反映を考慮し考案した SDF-ISF を使用する.SDF と ISF は次のように定義される.

・SDF(Speaker's Document Frequency):対象ユーザの全テキストにおける単語の出現頻度に基づく重み.

$$SDF=\frac{\text{対象ユーザの文書群における対象単語が出現する文書の数}}{\text{対象ユーザの文書数}} \quad (3)$$

・ISF(Inverse Speaker Frequency):全ユーザにおける対象単語を使用したユーザの希少度に基づく重み.

SDF と ISF の算出方法は以下である.

$$ISF = \frac{\text{全ユーザ数}}{\text{対象単語を使用したユーザ数}} \quad (4)$$

これら 4 種類の重みを組み合わせた「TF-IDF」,「SDF-ISF」,「TF-IDF-SDF-ISF」の 3 種類の重み付けを用いて作成した 3 種類のユーザ ID ベクトルを学習に使用し3種類の返答生成モデルを作成する.以下,これらのモデルを「TF-IDF モデル」,「SDF-ISF モデル」,「TF-IDF-SDF-ISF モデル」と呼称する.

3. データセット

3.1. 収集方法

データセットには,Twitter API を使用し収集したユーザ固有の ID である「ユーザ ID」,そのユーザによる「リプライ」,リプライ元の「ツイート」を使用する.以降,リプライを返答データ,ツイートを発話データとして扱う.

3.2. 返答生成モデル学習用データセット

返答生成モデルの学習用データセットとして,データ数300万のデータセットを作成した.このデータセットを使用し,ユーザ ID ベクトルの重み付けなしで学習させたモデルを「ベースラインモデル」と呼称する.

Improvement of style reflectivity of reply generation model by weighting user vectors with word preference,

†Itsuki Oishi, Masayasu Atsumi,
Information system sci., Graduate School of Sci. and Eng., Soka University†

3.3. テキスト分類器学習用データセット

収集したデータの中から,異なる顕著なスタイルを有する返答者5名を評価対象者として,評価に使用するためのデータを選別した. 評価対象者のデータの内 8 割のデータをテキスト分類器の学習用データセット,2 割のデータをテキスト分類器の検証用データセットとする.

4. 評価指標

評価指標は次の 3 つである.

・スタイル反映性：ID で指定された返答者のスタイルをどの程度反映できているか.
・文法正確性：日本語として適切な文法か.
・返答適切性：発話に対する返答として適切か.

次節より,以上 3 点の評価項目について,評価方法を述べる.

4.1. スタイル反映性

本評価では,返答者のユーザ ID をラベルとしてテキストをクラス分類する「返答者推定」を行う.テキスト分類器の学習には,評価のために収集したスタイルが顕著な返答者のユーザ ID と返答データを用いる.テキスト分類器は「BERT 日本語 Pretrained モデル[2]」をファインチューニングし作成した.

4.2. 文法正確性・返答適切性

被験者の主観による5段階評価を行う.各返答に対し,①適切,②概ね適切,③どちらともいえない,④やや不適切,⑤不適切の5項目のいずれかを選択してもらう.評価結果は,「適切」を 5,「概ね適切」を 4,「どちらともいえない」を 3,「やや不適切」を 2,「不適切」を 1 とし,被験者 4 名(本学学生)の評価の平均値を算出した.

5. 実験

5.1. テストデータ

返答生成モデルに,ランダムに収集した発話 20 文とスタイルが顕著な返答者のユーザ ID5 つを組み合わせて入力することで,返答を 100 文生成する.この生成を,「ベースラインモデル」,「TF-IDF モデル」,「SDF-ISF モデル」,「TF-IDF-SDF-ISF モデル」に対して行ない,返答を合計 400 文生成する.そして,これらを評価対象データとし,スタイル反映性,文法正確性,返答適切性の 3 項目を評価する.

5.2. 結果

表 1 に評価結果を示す.

表 1 評価結果

	スタイル反映性	文法正確性	返答適切性
ベースラインモデル	66	4.45	2.91
TF-IDFモデル	77	4.35	2.81
SDF-ISFモデル	76	4.28	3.25
TF-IDF-SDF-ISFモデル	78	4.63	3.34

5.3. 分析

本節では,4 つのモデルの重み付けと評価結果の関係について述べる.

まず,TF-IDF モデルはベースラインモデルと比べ,スタイル反映性が大きく向上した反面,文法正確性と返答適切性が低下した.これは,学習の比重がスタイルに偏った結果ではないかと考えられる.

また, SDF-ISF モデルはベースラインモデルと比べて文法正確性が低下したものの,スタイル反映性と返答適切性は向上した.これは,SDF-ISF がスタイル制御と返答生成の両方に適したバランスのいい重みであることを示している.

さらに,TF-IDF-SDF-ISF モデルは,スタイル反映性,文法正確性,返答適切性の全てにおいてベースラインモデルを上回るとともに,全モデルで最高評価値を達成した.これは,TF-IDF と SDF-ISF 両方の重みを使用することで,スタイル制御と返答生成に最も適した重み付けとなっていることを示している.

最後に, 重み付けと評価結果の関係について改めてまとめる.ベースラインモデルと比べて,どのモデルもスタイル反映性は向上した.一方,TF-IDF モデルは文法正確性と返答適切性が低下したが,SDFISF モデルは返答適切性が,TF-IDF-SDF-ISF モデルは文法正確性と返答適切性が向上した.このことから,重み付けの方法次第で評価指標全てで高い評価を達成することは可能であると考えられる.

6. 結論

本研究では,スタイルを制御可能な返答生成モデルの構築,ユーザベクトルの重み付けによるスタイル反映性・返答適切性の向上,テキスト分類器による返答者推定を利用したスタイルの客観的評価に取り組んだ.提案モデルの中で最も高い性能を発揮したモデルは TF-IDF-SDF-ISF モデルである.この結果は,ユーザベクトルを「文書全体から見た単語の重要度」と「ユーザ間の比較から見た単語の重要度」によって重み付けることで,モデルのスタイル反映性と返答適切性が向上することを示している.

参考文献

[1]tanreinama. "GitHub - tanreinama/gpt2-japanese: Japanese GPT2 Generation Model".2020.12.24. https://github.com/tanreinama/gpt2-japanese(参照 2022-1-25)

[2]黒橋 禎夫, 褚 晨翠, 村脇 有吾. "ku_bert_japanese - KUROHASHI-CHU-MURAWAKI LAB". 2019.4.1. https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese(参照 2022-1-25)