〈原著論文〉

チャットボットサービスのための GPT モデルを用いた NLU 訓練用データセット生成

片山義斗 *　中島潤 †

Generating NLU Training Datasets Using GPT Models

Yoshito KATAYAMA*　Jun NAKAJIMA†

要旨

フリーワード入力に対応したチャットボットを特定の組織で稼働させる際，その組織内で使用されている用語や言い回しに対応した訓練用データセットを用意して NLU モデルの訓練を行う方法がある。本研究では，OpenAI の GPT モデルである gpt-3.5-turbo-16k と gpt-4 を用いて訓練用データセットの自動生成を行い，両モデルのチャットボット用データセット生成能力の比較を行うとともに，訓練した NLU モデルの精度を計測し有効性を検証した。

Abstract

When using a chatbot that supports free-text input within a specific organization, one approach is to prepare a training dataset that matches the terminology and expressions used in that organization to train the Natural Language Understanding (NLU) model. In this study, in this study, we conducted three main activities: (1) automatically generated training datasets using OpenAI's GPT models, gpt-3.5-turbo-16k and gpt-4, (2) compared the dataset generation capabilities of both models for chatbots, and (3) measured the accuracy of the trained NLU model to verify its effectiveness.

キーワード

ChatGPT　チャットボット（Chatbot）　自然言語理解（NLU, Natural Language Understanding）　gpt-3.5-turbo-16k　gpt-4

* 北海道情報大学大学院経営情報学研究科修士課程, Master’s degree course, Department of Business and Information Systems, HIU

† 北海道情報大学経営情報学部システム情報学科教授, Professor, Department of Business and Information Systems, HIU

1. はじめに

企業や学校において顧客や従業員，学生や保護者などからの問い合わせに回答する方法として，メールや電話など対面で対応する方法や，専用の Q&A ページや Web サイトなどを用意して人が直接対応しない方法などがある。近年では，人が介在せずにチャット形式で質問と対応を行うチャットボットによる問い合わせ対応を導入する事例が多くみられる。コロナ禍や人手不足といった要因で，対面で問い合わせ対応を行うことのコストやリスクが高くなっており，チャットボットのように非対面で人が介在せずに質問対応ができるシステムは一層重要視されている。また，役所などでは窓口に向かう前に目的の窓口の所在をチャットボットで確認し，免許証や印鑑など事前に持参する必要のあるものを確認しておくことで，窓口で双方がスムーズな対応を行えるといった利点もあり，問い合わせ窓口担当者の負担軽減も期待できる。

図 1　北海道チャットボットサービス

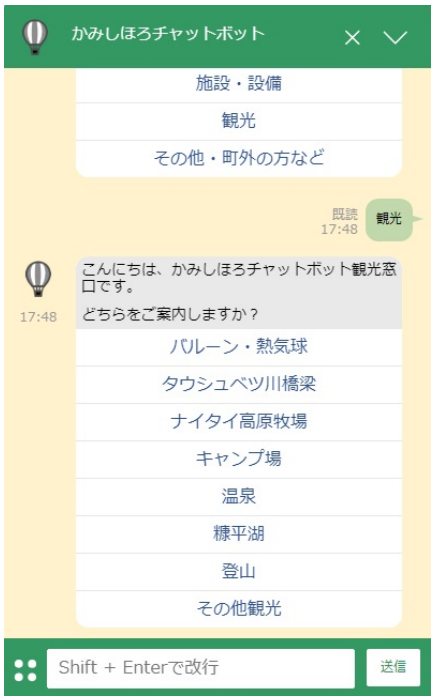

図 2　かみしほろチャットボット

自治体におけるチャットボット導入事例として北海道チャットボットサービス（https://chabo.pref.hokkaido.lg.jp/cb/gI9e-548N/main）（図 1）を挙げる。このサービスでは北海道のコロナウイルスやワクチン，感染時の対応といった情報の問い合わせを行うことができる。また，かみしほろチャットボット（https://www.kamishihoro.jp/page/00000321）（図 2）では行政サービスの手続きや町内施設についての問い合わせを行うことができる。

上記の例のように，チャットボットには選択肢から質問文を選択していく操作方法と，任意の質問文を入力し回答を得るフリーワード入力がある。例えば想定される質問文「落とし物をした」に対応する回答文「落とし物センターまでお越しください」がセットになったシナリオを用意するチャットボットシステムにおいて，選択肢から質問文を選んでいく方法では目的の質問にたどり着くまでに何度も操作が必要

となる。一方で，フリーワード入力に対応する方法では少ないやり取りで質問に返答できることが期待できる。しかし，質問者からは「落とし物はどこにありますか？」「財布を無くした」「鍵はどこ？」のように様々な言い回しで質問されるため，これらを「落とし物をした」という 1 つの意図に解釈する必要がある。そのため，フリーワード入力に対応したチャットボットでは一般的に Natural Language Understanding（以下 NLU と略す）エンジンによる質問の意図解釈を行う必要がある。しかし，NLU エンジンを利用するには機械学習により訓練した NLU モデルが必要となり，その訓練には訓練用のデータセットを用意する必要がある。チャットボット用に一般公開された汎用的なデータセットも存在するが，特定の組織向けのチャットボットサービスを考えた際，サービスごとに独自のデータセットを生成する必要がある。例として「（教室名）はどこ」「（学内 Web サービス）へのリンクを教えて」のような，チャットボットサービスを導入したい特定の組織限定の質問に対しては，独自に様々な言い回しの表記ゆれ質問文を用意する必要があるが，1 つの質問文に対して何百，何千もの表記ゆれ質問文を用意することは容易ではない。

本研究では，フリーワード入力に対応したチャットボットサービスで用いる NLU モデルの訓練のために，OpenAI（https://openai.com/blog/openai-api）の GPT モデルを用いて訓練用データセットを生成することを提案し，その有効性について検証した。さらに，訓練用データセット生成に使用した GPT モデルである gpt-3.5-turbo-16k と gpt-4 が様々な言い回しや表現の表記ゆれ質問文を生成しているか確認するため，生成した訓練用データセット内の表記ゆれ質問文間の類似度や，木村・高須ほか（2013）の編集距離を参考にレーベンシュタイン距離の計算を行い，両モデルのフリーワードに対応したチャットボットサービスの構築に使用する NLU モデル訓練用データセット生成という観点から性能や特徴を比較し評価を行った。これにより，訓練用データセット生成のために低コストの gpt-3.5-turbo-16k でも必要十分か gpt-4 を使う必要があるか比較実験を行った。

2.　OpenAI の GPT モデルを用いた訓練用データセット生成の提案

本研究において，NLU モデルの訓練に必要となるデータセットは質問文とその意図の 2 組のデータで構成される。質問者が「本を借りたいです」「本借りたい」「本借り方」のように様々な言い回しで質問しても「本を借りたい」という 1 つの意図に解釈する必要があるため，上記 2 組のデータを多数用意する必要がある。しかし，1 つの質問意図に対して多くの表記ゆれ質問文を用意することは容易ではないため，OpenAI の GPT モデルを用いて多数の表記ゆれ質問文を自動生成し，訓練用データセットの生成を行うことを提案する。

OpenAI が提供する GPT モデルに対して「”本を借りたい”という意図の表記ゆれ質問文を 100 生成してください」のように要求することで，100 の表記ゆれ質問文が得られる。これにより，手作業により表記ゆれ質問文を生成するよりも効率的に独自のデータセットを生成できることが期待できる（図 3）。

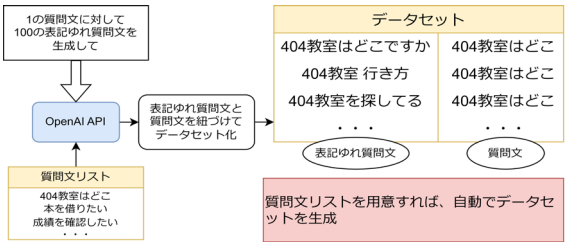

図 3　データセット出力方法の案

3.　訓練用データセット生成の実験

事前に用意した大学生活に関係する 50 の質問文に対して，それぞれ 100 の表記ゆれ質問文を出力するように要求し，同一のプロンプトによるデータセット出力結果の差について検証した（表 1）。また，両モデルによる 50 の質問文に対する表記ゆれ質問文出力数のグラフを出力した（図 4）。縦軸は表記ゆれ質問文の出力数を表し，横軸は総数 50 ある質問 ID を表す。なお，表記ゆれ質問文が重複した際や，意図していない出力形式やタイムアウトといったエラーが発生した際は，その質問文に対して再度出力要求を行った（図 5）。

表 1　データセット出力結果

GPT モデル	gpt-3.5-turbo-16k	gpt-4
表記ゆれ質問文出力数	3,608	4,731
平均表記ゆれ質問文出力数	72.16	94.62
標準偏差	41.84	7.88
重複回数	654	35
エラー回数	29	2
総再出力要求回数	683	37

gpt-3.5-turbo では GPT モデルへのプロンプトの文字数が不足するため，gpt-3.5-turbo よりも 4 倍のプロンプトを扱える gpt-3.5-turbo-16k を使用した。また，OpenAI の API で使用可能な GPT モデルの中で検証時に最も性能の高かった gpt-4 も使用し，モデル間のデータセット生成能力について比較を行った。

出力した表記ゆれ質問文数では，どちらも目標の 5,000 に到達しておらず，両モデル間で 1,123 の差異が見られた。gpt-3.5-turbo-16k の場合，1 つの質問文に対する表記ゆれ質問文出力数のばらつき具合（標準偏差）が大きく，要求数に対して過不足の出力が見られた。また，重複とエラーによる再出力回数も gpt-4 に比べ大きい値となった。結果として 3,608 の表記ゆれ質問文が得られた。

一方で gpt-4 の場合は，gpt-3.5-turbo-16k に比べて目標の 5,000 に近い出力数となった。1 つの質問文に対する表記ゆれ質問文数のばらつき具合，重複とエラーによる再出力回数も gpt-3.5-turbo-16k と比較し少ないことが確認された。

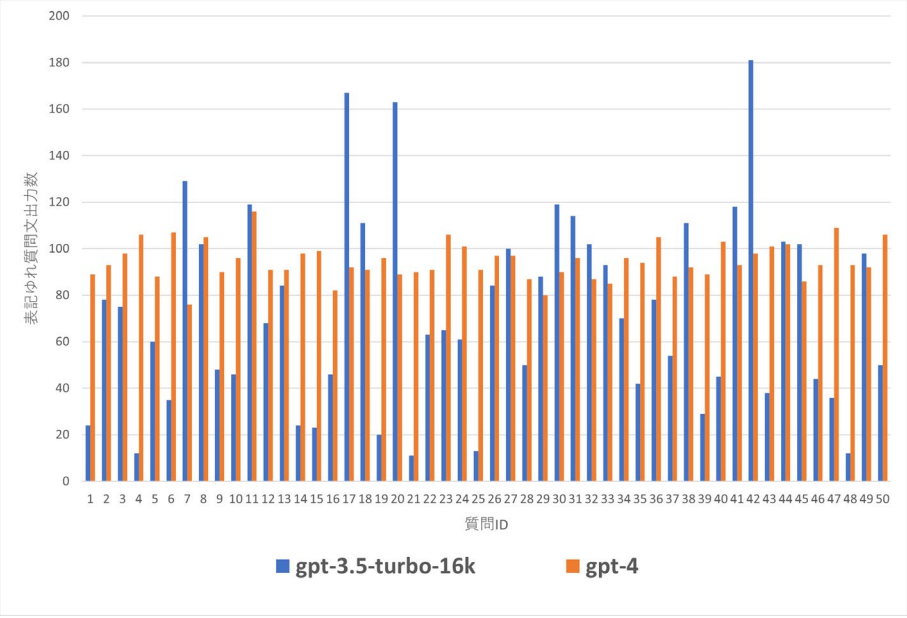

図 4　質問文に対する表記ゆれ質問文出力数

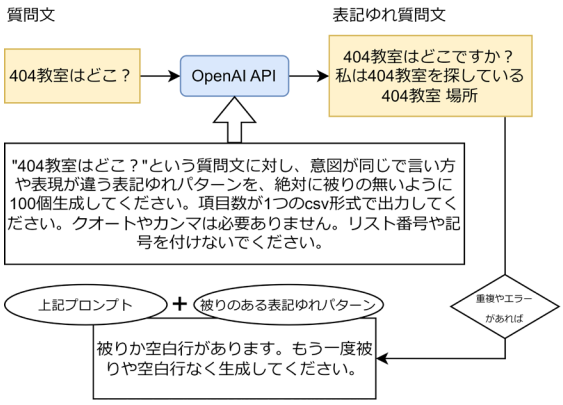

図 5　表記ゆれ質問文の出力

4.　類似度とレーベンシュタイン距離による GPT モデルの性能比較

GPT モデルにより出力したデータセット内の表記ゆれ質問文の中には「財布忘れた」「財布忘れた。」「財布忘れました」のように似たような言い回しが多く登場する。忘れ物に対応するシナリオでは，上記の言い回しの他にも「鍵を忘れた」「印鑑をなくした」のように，多様性のある表記ゆれ質問文を用意必要がある。そのため，生成した訓練用データセット内の表記ゆれ質問文間の類似度やレーベンシュタイン距離を計算することで，多様性のある表記ゆれ質問文を出力できる GPT モデルを選択することを提案し，類似度やレーベンシュタイン距離の計算を行った（図 6）。

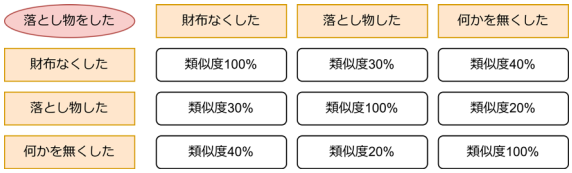

図 6　類似度計算の例

生成した表記ゆれ質問文間の類似度を，文字列同士の類似度や違いを表現することのできる，コサイン類似度・レーベンシュタイン距離・N グラム類似度（本研究では 2 文字）により比較した。コサイン類似度は難波（2020）を，N グラム類似度は村田・黒岩ほか（2002）を参考にした。

1 つの質問文ごとに複数の表記ゆれ質問文が存在するため，それらを全ての組み合わせで計算し平均を求めた（表 2）。

表 2　類似度と距離の計算結果

	gpt-3.5-turbo-16k	gpt-4
Mean Cosine Sim	0.0065	0.0013
Mean N-gram Sim	0.25	0.17
Mean Levenshtein Dist	10.95	13.91

まず，コサイン類似度の平均は gpt-3.5-turbo-16k の方が gpt-4 に比べて約 5 倍，N グラム類似度の平均では約 1.5 倍となった。また，レーベンシュタイン距離の平均では gpt-4 の方が gpt-3.5-turbo-16k に比べて 3 大きいことを確認した。gpt-4 の方が gpt-3.5-turbo-16k に比べて，文字列間の類似度が低くレーベンシュタイン距離が長いという結果が得られた。

表記ゆれ質問文同士の類似度とレーベンシュタイン距離をヒートマップで可視化した。縦軸と横軸は表記ゆれ質問文で，白い斜線は同じ表記ゆれ質問文同士の比較であり類似度やレーベンシュタイン距離が全く同じであることを表す。コサイン類似度と N グラム類似度はグラフのドットが明るいほど類似度が高く，レーベンシュタイン距離では青色が濃いほど距離が離れていることを表している。以下は，gpt-3.5-turbo-16k と gpt-4 で大きな差が見られた質問を取り上げ，両モデルの特徴について考察する。

質問 10 は「サークル・部活を作りたい」という質問となっており，コサイン類似度のヒートマップを視覚的に比較すると gpt-3.5-turbo-16k が gpt-4 よりも赤い点が多く見られる（図 7）。出力された表記ゆれ質問文を確認すると，gpt-3.5-turbo-16k では 45

の表記ゆれ質問文が「サークルと部活」「サークル・クラブ」「サークルや部活」の 3 つから始まる文章のみとなっている。一方 gpt-4 は上記 3 つ以外にも「部活動」「全クラブ」「全サークル」「部活及びクラブ」といった様々なパターンが見られる（表 3）。

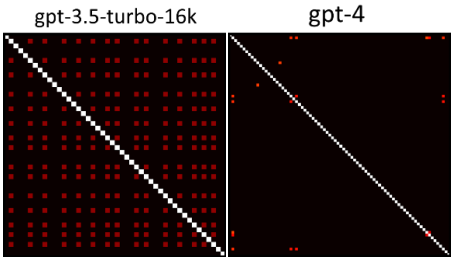

図 7　質問 10 のコサイン類似度ヒートマップ

表 3　出力したデータセット（質問 10 を一部抜粋）

gpt-3.5-turbo-16k	gpt-4
サークルと部活の一覧を見たい	サークルと部活の一覧を表示してください
サークル・クラブ一覧を見たい	全ての部活・サークルのリストが見たい
サークルや部活の一覧を知りたい	あなたは全てのクラブやサークルの一覧を持っていますか
サークルと部活のリストを欲しい	全ての部活とサークルの詳細を参照したい
サークルや部活のリストが見たい	全クラブとサークルの一覧を教えてくれる？
サークル・部活の一覧がほしい	部活動とサークルのリストを提供していただけますか
サークルと部活の一覧を教えてほしい	部活動やサークルの全一覧を見せていただけませんか？
サークルや部活のリストを教えてほしい	全ての部活とサークルのリストを見せてほしい
サークル・クラブの一覧が見たい	部活やサークルのリストを提供してくれませんか
サークルと部活のリストを知りたい	部活及びサークルの一覧表を教えて下さい
サークルや部活の一覧がほしい	部活・サークルの全てのメンバーを教えてもらえますか
サークルと部活の一覧を教えてください	部活とサークルの一覧が見たい
サークル・部活のリストを見たい	全クラブとサークルの情報を確認したい

質問 40 は「本の返却方法が知りたい」という質問となっており，レーベンシュタイン距離のヒートマップを視覚的に比較すると gpt-4 が gpt-3.5-turbo-16k より青色が濃く見られる（図 8）。出力された表記ゆれ質問文を確認すると，gpt-3.5-turbo-16k では「本の返却方法を教えてください」「本の返却方法を教えてくださいますか」「本の返却方法を教えていただけませんか」や「本の返却の方法について教えてください」「本の返却の方法について教えてくださいますか」「本の返却の方法について教えていただけませんか」というように，語尾だけを少しずつ変えたパターンを複数出力しており，グラフに市松模様が確認できる。一方で gpt-4 は書き出しや内容が不規則であり，距離が離れていることが確認できる（表 4）。

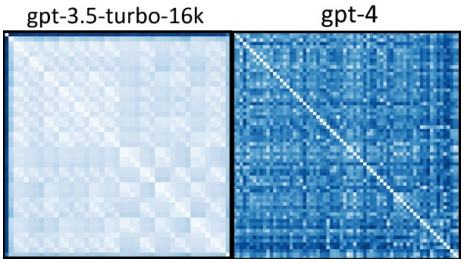

図 8　質問 40 のレーベンシュタイン距離ヒートマップ

表 4　出力したデータセット（質問 40 を一部抜粋）

gpt-3.5-turbo-16k	gpt-4
本の返却方法を教えてください	返却する手続きを教えてください
本の返却方法を教えてくださいますか	どうすれば本を返せますか
本の返却方法を教えていただけませんか	図書を返す方法について詳しく教えてください
本の返却の方法について教えてください	本の提出方法を教えてください
本の返却の方法について教えてくださいますか	書籍を返す流れは何ですか
本の返却の方法について教えていただけませんか	どう進めば本が返せますか？
本の返却の手続き方法を教えてください	戻すための方法を知りたいです
本の返却の手続き方法を教えてくださいますか	どういう手段で本を返せますか
本の返却の手続き方法を教えていただけませんか	本の返還方法を教わりたいです

質問 6 は「キャンパス，フロアマップが見たい」という質問となっており，N グラム類似度のヒートマップを視覚的に比較すると gpt-3.5-turbo-16k が gpt-4 よりも明るい点が多く見られる（図 9）。

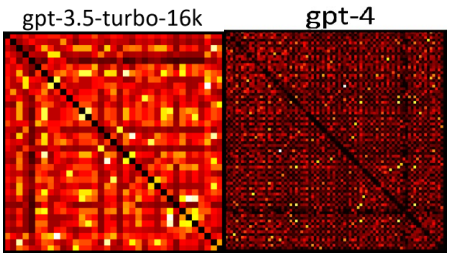

図 9 質問 6 の N グラム類似度ヒートマップ

出力された表記ゆれ質問文を確認すると，gpt-3.5-turbo-16k では 45 の表記ゆれ質問文が「フロアマップ」「キャンパスの」から始まる文章のみとなっている。一方 gpt-4 は上記 3 つ以外にも「キャンパスマップ」「フロア」「建物の」「詳細な」「どうやって」といった様々なパターンが見られる（表 5）。

表 5 出力したデータセット（質問 6 を一部抜粋）

gpt-3.5-turbo-16k	gpt-4
フロアマップをご覧になりたいですか。	建物のマップを知りたい
キャンパスのフロアマップはお持ちですか。	どのようにキャンパスレイアウトを見ますか
フロアマップを見せていただけますでしょうか。	フロアの間取り図が確認したい
フロアマップがほしいです。	キャンパスの平面図を見せてほしい
キャンパスのマップを見たいです。	フロアのマップを教えてください
フロアマップを見ることができますか。	キャンパスの間取り図が見たい
フロアマップを教えていただけますか。	詳細なフロアマップが見たい
フロアマップを見せてほしいです。	キャンパスのマップを提供してくれますか
フロアマップを教えてほしいです。	フロアの平面図が欲しいです
キャンパスのフロアマップを教えてくれますか。	キャンパスマップを確認したいです
フロアマップを見れる場所はありますか。	キャンパスの詳細図を教えてください
フロアマップを見せていただけると助かります。	フロアの詳細図を見る方法はありますか
フロアマップを提供していただけますか。	キャンパスの建築計画図を閲覧したい

以上のことから, gpt-3.5-turbo-16k よりも gpt-4 の方が表記ゆれ質問文を様々な言い回しで生成できることを確認した。

5. NLU モデルの訓練

両モデルにより得られたデータセットを元に，OSS の NLU エンジンである COTOBA Agent dialogue engine（https://github.com/cotobadesign/cotoba-agent-oss）を用いて，NLU モデルの訓練と F1 スコアの測定を行った。訓練時のハイパーパラメータは初期値を用いて，nbatch:32，niter:3，num_warmup_steps:0, num_training_steps:3 という条件で訓練を行った。ハイパーパラメータで一般的に用いられる用語に置き換えると，nbatch はバッチサイズ，niter はエポック数を意味する。num_warmup_steps は huggingface における scheduler に使用するハイパーパラメータであり，0 を指定すると訓練開始時から最大の学習率が適用される。num_training_steps は訓練を行うステップ数を指定し，3 を指定するとモデルが 3 ステップで訓練を完了する事を意味する。

まず，可能な限り同条件で NLU モデルの比較を行うため, gpt-4 の表記ゆれ質問文の出力数を gpt-3.5-turbo-16k に合わせ，プロンプトに 50 の質問文に対してそれぞれ 80 の表記ゆれ質問文を出力するように要求した。その結果 3,872 の表記ゆれ質問文が出力され，gpt-3.5-turbo-16k との差は 264 となった。

両モデルが出力したそれぞれのデータセットを，トレーニングデータ・検証データ・テストデータの 3 つに，7:2:1 という機械学習において一般的に使用されている比率で分割を行った。データセット出力時に拾いきれなかったエラーや空白行といった欠損値を取り除いた，最終的なデータセット分割後の表記ゆれ質問文の内訳を表 6 に示す。データセットの分割時は，トレーニングデータ・検証データ・テストデータ全てに, 50 種類の質問文に対応する表記ゆれ質問文が均等に振り分けられるように設計した。

表 6 データセットの分割後の内訳

モデル	gpt-3.5-turbo-16k	gpt-4
表記ゆれ質問文出力数	3,608	3,872
トレーニングデータ数	2,504	2,676
検証データ数	720	776
テストデータ数	381	407
欠損データ数	3	13

COTOBA Agent dialogue engine の NLU エンジンでは，検証データを用いて最小の損失を持つモデルの選定を行う。また，テストデータを用いて F1 スコアを計算することができ, F1 スコアが 1 に近いほど訓練した NLU モデルが質問意図を正しく意図解釈できることを確認できる。両モデルで出力したデータセットを用いて訓練されたモデルの F1 スコアを表 7 に示す。

Macro の F1 スコアでは，各質問（クラス）の F1 スコアの平均を，Micro の F1 スコアでは，単純にテストデータ全ての正解率を確認できる。結果として，Macro の F1 スコアでは gpt-4 の方が 0.043 高く，Micro の F1 スコアでは gpt-3.5-turbo-16k の方が 0.05 高いことが確認された。

以上のことから，OpenAI API により出力されたデータセットを用いることで，約 80%の精度で意図解釈が可能な NLU モデルを訓練できることを確認した。

表 7　F1 スコアの比較

モデル	gpt-3.5-turbo-16k	gpt-4
表記ゆれ質問文数	3,608	3,872
Macro average F-measure	0.751	0.794
Micro average F-measure	0.915	0.865

6. おわりに

本研究では，OpenAI の GPT モデルを用いて，フリーワード入力に対応したチャットボット用の NLU モデル訓練用データセットの生成を行うことを提案し，類似度やレーベンシュタイン距離から両モデルの性能を多様性のある表記ゆれ質問文の出力という観点から比較を行い，最終的に F1 スコアが約 80%あることから実用できる可能性があることを確認した。

現在は，フリーワード入力において「履修手続きと卒業要件について知りたい」のように 2 つ以上の質問を同時に投げかけられたときに，どちらか片方の質問への回答しか行えないため，質問文から 1 つ 1 つの質問意図を分割・意図解釈し，回答を繋げる仕組みが必要である。また，本研究では質問の数を 50 で実験を行ったが，実際にチャットボットが稼働する現場では大量のシナリオが必要になると考えられる。そのため，対応できなかった質問（あらかじめ用意していないシナリオ）が投げられたときに，それに対応する質問（シナリオ）を都度増やしていく必要がある。また，GPT モデルの類似度を計算した結果，gpt-4 の方が gpt3.5-turbo-16k よりも優位性があるとしたが，実際に NLU モデルの訓練を行った結果は gpt3.5-turbo-16k に比べて gpt-4 は Macro の F1 スコアで 0.043 高く Micro では 0.05 低い結果となった。引き続き，gpt-4 用の訓練及びテストデータセットにおいてクラス不均衡の有無を確認し，シナリオ数を増やしたうえで実運用時に近い条件で検証を行いたい。

参考文献

COTOBA DESIGN「cotobadesign/cotoba-agent-oss」 https://github.com/cotobadesign/cotoba-agent-oss （2024 年 1 月 9 日アクセス）。

北海道感染症対策連絡本部指揮室「北海道新型コロナウイルス感染症　陽性者サポートサイト（札幌市、旭川市、函館市、小樽市にお住まいの方を除く）」 https://chabo.pref.hokkaido.lg.jp/cb/gI9e-548N/main （2024 年 1 月 30 日アクセス）。

上士幌町役場デジタル推進課「かみしほろチャットボット」 https://www.kamishihoro.jp/page/000003 21 （2024 年 1 月 30 日アクセス）。

木村・高須ほか(2013)「高頻度語を可変長索引語に用いる類似文字列検索手法の検討」『FIT2013（第 12 回情報科学技術フォーラム）』第 2 分冊，pp.151-152

村田・黒岩ほか(2002)「学生レポートの n-gram による類似度評価の検討」『FIT（情報科学技術フォーラム）』, pp.101-102。

難波英嗣(2020)「テキスト間の類似度の測定」『情報の科学と技術』第 70 巻 7 号，pp.373-375。

OpenAI「OpenAI API」 https://openai.com/blog/openai-api （2024 年 1 月 9 日アクセス）。