GPT を用いた標準語から方言への翻訳

山崎祐 [1]　坂野遼平 [1]

[1] 工学院大学 情報学部 情報通信工学科
j020297@ns.kogakuin.ac.jp

概要

現在日本では方言は当該地域における日常会話等にとどまらず，様々な形で活用されている．その方言を支える職業として方言指導講師がある．しかし若者の方言離れなどによる問題から方言指導講師の減少が懸念されている．指導できる人が減少すると舞台の原稿作成や方言の翻訳などが困難になってしまう．そこで本研究では，GPT を利用した標準語から津軽弁へのテキスト翻訳を行う．具体的には標準語と津軽弁の対訳データを Fine-tuning し，津軽弁への翻訳を指示する prompt を与える．実験の結果，学習前よりも Fine-tuning をした後のほうが BLEU スコアが高くなることが分かった．

1　はじめに

方言とはある言語が地域によって別々な発達をし，文法・イントネーション・語彙などの上で相違のあるいくつかの言語圏に分かれたとみなされた時のそれぞれの地域の言語体系のことである．東條操の方言区画論 [1] では，日本語の方言は本土方言と琉球方言に大別される．本土方言は東部方言，西部方言，九州方言の 3 つに分類される．これらはさらに，北海道で話される北海道方言や，鹿児島県や宮崎県で話される薩隅方言といった方言で細分化されている．また琉球方言は沖縄や奄美大島，宮古島で話される．本土方言は 13，琉球方言は 3 の計 16 の方言が日本には存在する．日本では異なる各地方独特の語彙や言い回しあるいはイントネーションや発音の違い，いわゆるなまりを指す場合が多い．

現在日本では方言は当該地域における日常会話等にとどまらず，様々な形で活用されている．例として地域資源として観光誘致に活用されるケースがある．具体的には空港や駅など公共交通機関や観光客が集まる土産物屋には歓迎を表すあいさつ方言が使われている．また劇団四季のライオンキングの舞台では地域講演に合わせ，セリフをそれぞれの土地に合った方言に変えるという演出がある [2]．このように観光誘致や多様な作品表現には方言が欠かせない．それを支える職業として方言指導講師がある．方言指導講師とはドラマや映画，舞台など俳優に方言を指導する専門家である．俳優が方言を習得できるように，練習の原稿を作成したり，方言のイントネーションや発音を指導する．このような専門家によって多様な方言の活用が支えられている．

しかし，方言指導講師の減少が懸念されている．一因として方言の標準語化が進んでいることがあげられる．これは，標準語に触れる機会が増えることで，方言に替わって標準語が多く用いられるようになる現象を指す．背景として 2011 年の国立国語研究所の調査 [3] によると，戦前以来の標準語教育や都市部への就職，マスメディアの発達が挙げられている．こういった背景に伴い，若者が地方から都市に移り，方言指導講師の担い手の減少が懸念される．それにより方言を指導できる人が減少すると舞台の原稿作成やイントネーションの指導が困難になってしまう．

本研究では津軽弁に着目し，Generative Pre-trained Transformer（GPT）を利用した機械翻訳でテキストベースの標準語から津軽弁への翻訳を行う．津軽弁とは青森県津軽地方で話される日本語の方言である．青森県内の方言はこの津軽弁と南部地方の南部方言に大きく分かれる．本研究では前者の津軽弁を扱う．研究目的は GPT を用いた言語モデルで原稿を作成することで方言指導講師の負担を軽減することである．

2　関連研究

今井の研究ではディープラーニングフレームワーク Chainer を用いて津軽弁と共通語双方向の音声・文字情報変換システムの開発を行った [4]．津軽弁テキストを入力として形態素解析を行うシステム

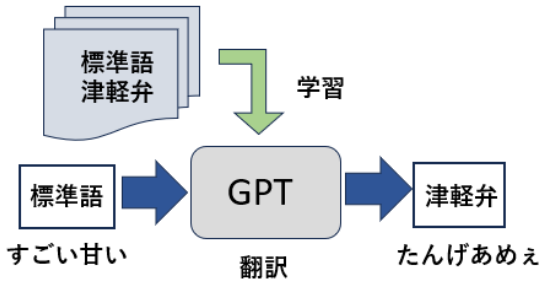

図 1　提案システム

と，人工知能により翻訳を行うシステムにより構成されている．形態素解析には京都大学情報学研究科と NTT コミュニケーション科学基礎研究所による共同研究ユニットプロジェクトを通して開発されたオープンソース MeCab (和布蕪) を使用している．形態素解析エンジンは独自のライブラリを追加できるため，方言情報を用いたライブラリを作成することができる．今井の研究では MeCab を使用して，津軽弁ライブラリを作成し形態素解析を行った．しかし分割精度は 62 %だったため，正しく分割できない例があった．本研究では GPT を用いることで標準語から津軽弁への翻訳を行い，精度を検証する．

尾崎らの研究では大規模言語モデルである GPT-3 の反論文生成品質を人手により評価・検証することに加え，収集した反論とそれを元に生成した生成反論の文章一致率を算出・評価を行った．反論生成は GPT-3 の prompt に議題と立論とそれに対する反論の回答例をあたえ，それをもとに反論を出力するようにした．この研究では prompt のクエリ部（「反論を生成せよ」という指示）が反論文生成に大きく寄与していることが分かった [5].

3　提案手法

図 1 に提案システムの動作の流れを示す．

本研究では GPT に標準語から津軽弁の翻訳を行わせる．初めに GPT に方言データを学習させる．方言データは標準語と津軽弁の対訳のデータである．具体的な候補として，国立国語研究所の全国方言データ [6] や，先行研究の今井の研究 [4] で用いられている標準語と津軽弁の対訳文章などが挙げられる．前者の方言データは標準語と津軽弁単語単位の対訳に対して，後者は文章データとなっている．本研究では後者の文章データを学習させる．なぜなら GPT は transformer をもとに作られている．transformer は前後の文脈から，次にどの単語が来るべきかを予測するように訓練する．そのため単語ではわからない前後関係を，文章で学習させることができる．

学習を行った GPT に入力として標準語を与える．ここで GPT に対しての prompt を設定する．prompt は「(標準語文) を津軽弁に翻訳してください.」とした．方言に特化した GPT がその入力を翻訳する．そうすることで津軽弁に翻訳されたものが出力される．

3.1　GPT

GPT とは，自然言語の生成や応答，会話のモデリングに特化して開発された言語モデルである．これらのモデルは訓練データとしてニュース記事や Web ページ，書籍，会話ログなど莫大なテキストデータが使用されている．GPT をある目的に特化させるための手段として Fine-tuning と prompt が挙げられる．

3.1.1　Fine-tuning

Fine-tuning とは追加のデータを用意してモデル自体をさらに学習させる方法である．GPT は事前学習された汎用モデルであり，特定のタスクに対して最適化されていないことがある．そこで自分の行いたいタスクに特化させるために，追加のデータを学習させることでモデルのパラメータを調整する．OpenAI によると Fine-tuning をすることで prompt よりも高品質な応答が可能なことや prompt の短縮による処理時間の短縮が期待できる [7].

3.1.2　prompt

prompt とは言語モデルに対しての指示や命令のことである．prompt を適切に設定することで，言語モデルの出力が良質なものになる．基本的な prompt のパターンとして zero-shot prompting および few-shot prompting がある．前者の「zero-shot prompting」は例を与えず，直接質問を与える prompt の型である．すでに多くのデータが学習されているモデルであれば正しい回答が得られるが，間違った回答が出力される場合がある．一方後者の「few-shot prompting」ははじめにいくつか指示とその回答のセットを与え，学習させる prompt の型である．このセットの数が多ければ，適切な回答が得られる．

4　評価手法

本研究では翻訳した生成文と正解訳を評価するために BLEU スコアと人手評価を用いる．

4.1　BLEU スコア

BLEU スコアとは翻訳したフレーズと正解訳に含まれているフレーズが比較され，そのフレーズの一致数をカウントするスコアである．機械翻訳の性能を評価するための自動的な尺度で，0〜1 の範囲であらわされ，正解訳に近いほど数値が高くなる．このフレーズは N-gram で分割したテキストである．N-gram とは任意の文字列や文書を連続した n 個の文字で分割するテキスト分割方法である．例えば「今日は暑いですね」を単語単位 1-gram で分割すると，「今日，は，暑い，です，ね」となる．このように分割したフレーズ同士を比較し，一致している数で評価する．算出式は以下のようになっている

$$BLEU = e^{(1-r/c)} exp(\sum_{i=1}^{n} w_n log p_n)$$

r は人間が翻訳した文章の長さを表し，正解訳の意味で reference の頭文字をとっている．c は機械が翻訳した文章の長さを表し，機械が翻訳した文の候補の意味で candidates の頭文字をとっている．

$w_n = 1/N$

$p_n = \frac{\Sigma_{C \in Candidates} \Sigma_{n-gram \in C} Count_{clip}(n-gram)}{\Sigma_{C' \in Candidates} \Sigma_{n-gram' \in C'} Count(n-gram')}$

と定義される．一般的に N=4 と定義されることが多い．

$count_{clip}(n-gram)$ は n-gram のときに，機械が翻訳した文章（candidates）と人間が翻訳した文章（candidates）の一致している数である [8].

BLEU スコアには弱点がある．それは文章の意味を考慮していないことである．BLEU スコアは字面しか見ていないため，同義語でも完全に文字が一致しなければ，評価されない．そのため翻訳を正しく評価するためには BLEU スコアのみでは不十分である．そこで人手評価で文章の意味や表現の評価を補う．

4.2　人手評価

人手評価は評価型国際ワークショップ IWSLT（International Workshop on Spoken Language Translation[9] で採用された基準を用いる．今回は流暢さに着目する．流暢さとはある言語として自然な表現であるかを示す．つまり津軽弁の文章として自然な文章であるかどうかを見てもらう．今回は流暢さを 1 5 の 5 段階で評価してもらう．

表 1　gpt-3.5-turbo と提案手法の BLEU スコア

モデル	BLEU
gpt-3.5-turbo	0.074
提案手法	0.159

5　実験

今回使用したモデルは gpt-3.5-turbo である．方言データは今井の研究 [4] や津軽弁をまとめたホームページ [10] から収集した文章 207 文を使用した．そのデータを学習データとテストデータで 8:2 に分けた．学習データ 165 文を gpt-3.5-turbo に Fine-tuning をした．そのモデルには「(標準語文) を津軽弁に翻訳してください.」という prompt を指定した．そうすることで GPT で翻訳をすることができる．そしてテストデータ 42 文を用いて翻訳した生成文と正解訳を BLEU スコアで評価した．BLEU スコアは N-gram で分割するため単語で分かれている必要がある．そのため評価をする前に，生成文を津軽弁の辞書を用いた janome で形態素解析を行った．さらに人手評価を行った．津軽地方出身の 4 人の評価者に行ってもらった．Fine-tuning する前とした後の 42 文をそれぞれ 1〜5 で評価してもらい，それを平均したもので比較した．

5.1　評価結果

Fine-tuning する前の gpt-3.5-turbo と提案手法のモデルで BLEU スコアを用いて評価した結果を表 1 に示す．また提案手法の成功例と失敗例の出力結果の例を表 2 と表 3 に示す．Fine-tuning をする前の BLEU スコアと比べ，Fine-tuning を行った提案手法は倍以上の値となった．一般的に Fine-tuning を行うことで，あるタスクの性能が上がる．この実験により，方言の翻訳でも BLEU スコアが上がることが分かった．しかし BLEU スコアは 0.4 以上の値になると高品質な翻訳といわれているため，高品質な翻訳とは言えない．

次に人手評価の結果を表 4 に示す．表 4 は 1〜5 段階で評価してもらったものを平均し，Fine-tuning 前の 42 文の平均値と，Fine-tuning 後の平均値で比較した結果である．Fine-tuning 後のほうが平均値は低くなっている．42 文のうち，Fine-tuning 前のほうが高い評価となったものは 25 文，Fine-tuning 後のほう

表 2 出力結果（成功例）

	標準語	翻訳結果
正解訳	またね	へばねー
提案手法	またね	へばねー

表 3 出力結果（失敗例）

	標準語	翻訳結果
正解訳	このお風呂随分気持ちいいね	この風呂ずんぶあずましいな
提案手法	このお風呂随分気持ちいいね	このお風呂すげーたのしいな

表 4 人手評価の平均値

Fine-tuning 前	Fine-tuning 後
2.261	2.208

が高い評価となったものは 12 文あり，同一評価が 5 文であった．

5.2 考察

BLEU スコアに関しては，Fine-tuning 前よりも後のほうがスコアが高くなっていることが分かった．しかし人手評価での流暢さの評価は Fine-tuning 前よりも後のほうが低くなっていることが分かった．なぜなら BLEU スコアはフレーズの一致率を見るため，意味については考慮されない．そのため Fine-tuning したことで単語は方言らしくなったが，実際津軽弁の文章としての意味はおかしくなってしまった．

6 おわりに

方言を支える職業として方言指導講師がある．しかし，言葉の標準化により，方言話者は減り，さらに彼ら/彼女らの高齢化も問題となっている．そこで大規模言語モデルである GPT を利用して，標準語から方言の翻訳を行った．GPT を方言に特化させるモデルにするために，Fine-tuning を行った．結果としては Fine-tuning を行うことで標準語から津軽弁への翻訳の BLEU スコアが上がることが確認できた．しかし一般に高品質と言われる BLEU スコアには達しておらず，十分な翻訳精度とは言えない．また人手評価により Fine-tuning する前よりもした後のほうが流暢さが低くなっていることが分かった．

今後の課題として翻訳精度と形態素解析の精度を高めるためにさらに津軽弁文章を収集する必要がある．また方言の機械翻訳の prompt による効果で zero-shot-prompting と few-shot-prompting の精度比較を行っていきたい．

参考文献

[1] 東條操. 日本方言学. Technical report, 吉川弘文館, 1954.

[2] 劇団四季. https://www.shiki.jp/applause/lionking/learn_more/more_lk_2.html(accessed August. 1 2023).

[3] 木部暢子. 方言の多様性から見る日本語の将来. 学術の動向, Vol. 16, No. 5, pp. 5 108–5 112, 2011.

[4] 今井雅. 特別講演「あなたの津軽弁を共通語に——弘大× ai ×津軽弁の取り組み——」. 日本放射線看護学会誌, Vol. 10, No. 1, pp. 9–12, 2022.

[5] 尾崎大晟, 中川智皓, 内藤昭一, 井之上直也, 山口健史. 大規模言語モデルが生成した反論文の品質評価. In **The 37th Annual Conference of the Japanese Society for Artificial Intelligence**, p. 4, 2023.

[6] 国立国語研究所. https://www2.ninjal.ac.jp/hogen/dp/fpjd/fpjd_index.html(accessed January. 4, 2024).

[7] OpenAI. https://platform.openai.com/docs/guides/fine-tuning(accessed December. 20, 2023).

[8] Kishore Papineni, Salim Roukos Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**, pp. 313–315, 2002.

[9] Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, Jun'ichiTsujii. Overview of the iwslt04 evaluation campaign. In **IWSLT 2004**, p. 3, 2004.

[10] Moritaka Ogasawara. https://www2d.biglobe.ne.jp/~oga/tsugaru/ben.html(accessed January. 4, 2024).