大規模言語モデルを用いたカタカナ語の意味分類における出力傾向分析

小滝主紀 [1] 佐々木稔 [2]
[1] 茨城大学大学院 理工学研究科 [2] 茨城大学 工学部 情報工学科
{24nm724g, minoru.sasaki.01}@vc.ibaraki.ac.jp

概要

大規模言語モデル (LLM) を用いた外来語の意味推論において、和製英語の存在など元来の英単語の意味との違いがモデルの精度に影響を与える可能性がある。外来語の意味を正確に捉えるために、本研究では BCCWJ から抽出したデータで LLM を Fine-tuning し、現状のカタカナ語の意味推論の精度及び出力傾向を分析、精度向上への方策を探る。複数の実験の結果、Zero-shot learning では Fine-tuning の有効性が確認できなかったが、対照的に Few-shot learning では約 10 %の精度の向上が見られた。出力傾向分析によってモデルの苦手な語義や単語の傾向を得られたため、今後の研究への寄与を期待する。

1 はじめに

現在、自然言語処理分野において LLM に関連した研究が行われている。LLM を含む生成 AI の中でも認知率、使用率がともに高い ChatGPT は、英語データを主に利用し作成されており内部で占める日本語データの割合は低いと推測される。さらに日本語に含まれるカタカナ語には、英語を語源とする外来語や和製英語がある。これらは対応する英単語の元来の意味と異なる場合があり、日本語データが少ないことも相まって、文脈中の意味分類が正しく行われない可能性が大きい。そこで本稿では、国立国語研究所が提供する「現代日本語書き言葉均衡コーパス [BCCWJ(Balanced Corpus of Contemporary Written Japanese)]」[1] からカタカナ語を含む文章を抽出してデータセットを作成し、これを用いて OpenAI 社の提供する gpt-4o-mini-2024-07-18 モデルに Fine-tuning を行い、出力の傾向分析を行う。これによりモデルの苦手なカタカナ語、そして対象単語の持つ推測しにくい意味の傾向を分析し、語義曖昧性解消への貢献を期待する。

2 関連研究

機械学習に基づく語義曖昧性解消の研究には、大きく分けて教師あり学習手法と知識ベース手法の 2 つのアプローチがある。教師あり学習手法は、人間が対象となる多義語に正しい語義のラベルを付けたコーパスによって学習した語義分類モデルを用いて、曖昧な単語に対して適切な意味を分類する。近年の教師あり語義曖昧性解消モデルでは BERT[2] や RoBERTa[3] といった事前学習された言語モデルを用いて単語や文をベクトル化して語義の識別を行っている [4, 5]。知識ベース手法は語義ラベル付きコーパスを使わず、辞書やオントロジーといった外部知識を用いた分類手法である。単語の語義定義文をベクトル化して語義間の関係を学習する手法 [6] や単語間の類義関係を用いて効果的な語義のベクトルを求める手法 [7] などがある。また、最近では語義曖昧性解消において ChatGPT のように大規模言語モデルを用いた生成 AI を利用する試みも検討されている [8, 9]。評価データによる実験において高い性能を示す結果が得られているが、最先端のモデルが達成するレベルにはまだ達していない。

3 実験

実験で作成したソースコードやデータセットの一部は Github で公開している。詳細な内容については対応するデータを参照していただきたい。

3.1 外来語抽出

本実験のデータセットを作成するにあたり、「現代日本語書き言葉均衡コーパス BCCWJ」を利用した。BCCWJ 内の短単位で区切られた SUW ファイルより、語種が外来語である単語を含む 32226 件の文を抽出した。全体に対する外来語の網羅率は約 88.7 %となった。対象単語を選別するため、外来語

を含む文章が 5 種類以上あることを条件にその単語を抽出すると、文数は 10750 文、単語数は 795 語となった。これらの単語を次節以降データセットを作成するために利用する初期データとした。

3.2 データセット作成

3.2.1 対象単語の選別

本節ではデータセット中に含める対象単語とそのデータの選別方法について説明する。前節において対象単語に条件を付けたのは、訓練データとして 4 文、テストデータとして 1 文を最低でもデータセットに含めることを想定したためである。以上より訓練データ中の語義に偏りがないように訓練データ中に含まれる対象単語の意味が最低でも 4 つ以上あること (4 文であれば各文に含まれる対象単語の意味がそれぞれ異なること) を条件に、データセットに含めるべきか否かの選別を手作業で行った。選別するにあたり、NTT ドコモが提供する辞書検索サービスである goo 辞書[1)]の原典であるデジタル大辞泉を対象単語の意味区分を参照するために利用した。

選別の方法としては、まず初期データの対象単語 795 語について一つずつ goo 辞書内で手作業で検索し、4 つ以上の意味区分があるものを選別する。この際、カタカナ語のみを利用するため英単語や他言語の冠詞 (ラなど) 等は除く。さらに BCCWJ から抽出した文章中の対象単語の意味がデジタル大辞泉内の意味区分に含まれていることを手作業で確認し、最終的にデジタル大辞泉内で 4 つ以上の意味区分を持ち、かつ BCCWJ 内に対応する語義が 5 つ以上存在する単語をデータセット中の対象単語とする。以上の方法で選別した結果、単語数は 40、文章数は 1143 文であった。

3.2.2 訓練、テストデータの作成

データセットを作成するために、選別された対象単語 40 語とそれらを含む BCCWJ から抽出した 1143 文を、デジタル大辞泉中の意味区分に照らし合わせ、それぞれの文中での語義を手作業でタグとして付与した。また、BCCWJ から抽出した文が短く意味をとることができないものは 0、他の単語との複合語で元の意味と異なるものには 1、文脈によって意味が複数受け取れるものには 2、デジタル大辞泉の意味区分に含まれないものを 3 をタグ付けした。これら 0〜3 にタグ付けされた文章は訓練、テストデータには含めず (246 文)、実際に利用したデータは 897 文となった。

このデータの中よりランダムな抽出を行い、実験の条件に応じて訓練データとテストデータに分割を行う。後述する実験 1、実験 7 では、テストデータを各 40 語× 3 文ずつランダムに抽出し (120 文)、それ以外の 777 文を訓練データに利用した。対して実験 2、実験 3、実験 4、実験 5、実験 6、実験 8 では、テストデータをデジタル大辞泉の意味区分中の語義に対応する文が一つしかない文 (BCCWJ 内の語義の出現頻度が 1 である文) で構成し (56 文)、それ以外の 841 文を訓練データの作成に利用した。これは zero-shot learning での Fine-tuning の効果を確認するためである。

3.3 Fine-tuning

3.2 節で抽出した各訓練データを OpenAI 社が提供している API によって学習させ、Fine-tuning を行う。Fine-tuning を行ったモデルは gpt-4o-mini-2024-07-18、変更できるハイパーパラメーターは全て auto を指定した。

3.4 語義の予測

各実験の語義の予測は Fine-tuning 前のモデルと Fine-tuning 後のモデルを利用して ChatGPT に回答を生成させ、得られた予測の正否を手作業で判定し、精度の比較及び出力の傾向を確認する。出力の正否の基準は、出力内容がテストデータにタグ付けしたデジタル大辞泉の意味区分中の語義とおよそ一致していることである。なお、プロンプト中に選択肢としてデジタル大辞泉の意味区分を与えるように記述しており[2)]、推測した語義と最も近い選択肢を出力する旨も記述しているため、基本的には選択肢の中から回答が出力されるように指定している。

3.5 各実験内容

これより実際に作成したデータセットを利用して実験を行う。実験は 8 つ行い、Fine-tuning は OpenAI 社の提供する API を利用して行った。OpenAI API の Fine-tuning で利用できるデータセットの形式は決まっており、その形式に沿った訓練データを作成し、gpt-4o-mini-2024-07-18 を Fine-tuning し、作成

1) https://dictionary.goo.ne.jp/

2) goo 辞書よりスクレイピングした意味区分を利用しプロンプトに渡す

されたモデルを利用して実験を行う。モデルの出力を得る際に指定したハイパーパラメーターは temperature のみであり、すべての出力で 0.9 に指定した。temperature は 0 を指定すると回答が一意に決まり、0 より大きい値を指定すると回答にバラつきが見えるようになり、多様性に富む出力が得られるようになるパラメーターである (0〜2.0 までの値を取る)。複数の実験の出力の傾向を分析することを目的としているため、回答が一意に留まらないようにするためこのように指定した。以下では各実験でのデータセット内部の情報と実験目的を明示する。3.5.1 節では実験 1、3.5.2 節では実験 2、実験 3、実験 4、実験 5、実験 6、3.5.3 節では実験 7、実験 8 について使用したデータセットの詳細を示す。

3.5.1 Few-shot learning

本実験では、Few-shot learning における Fine-tuning による出力精度の上昇または低下を確認することを目的としている。実験 1 におけるテストデータは、各対象単語からランダムに 3 文抽出した 120 文 (対象単語 40 語× 3 文) であり、訓練データはそれらを除く 777 文である。

3.5.2 Zero-shot learning

本節の実験では、Zero-shot learning における Fine-tuning の有効性を確認することを目的としている。なお本節の各実験で、テストデータはデジタル大辞泉の意味区分中の語義に対応する文が一つしかない 56 文 (BCCWJ 内の出現頻度が 1 である意味を持つ対象単語を含む文章) であり共通である。実験 2 ではテストデータに含まれない 841 文を訓練データに利用した。実験 3 では、テストデータに含まれる対象単語を含む文章を訓練データ 841 文の中から抽出して利用した (596 文)。実験 4 では、実験 3 の訓練データ 596 文をランダムに半分抽出し利用した (298 文)。実験 5 では、訓練データ 841 文の中から実験 3 で利用した 596 文を除いた 245 文を訓練データとした (訓練データ中に対象単語を含む文は存在しない)。実験 6 では実験 4 と同じように、実験 5 の訓練データをランダムに半分抽出し利用した (122 文)。

3.5.3 対象単語タグと Chain-of-Thought の利用

実験 7 では、入力として文と対象単語、さらに取りうる語義をデータセットに与え、対象単語の最適な意味を抽出する手法を取る ESC(Extractive Sense Comprehension)[10] を参考にし、実験 1 の訓練データ、テストデータ中の対象単語をすべて $< t >< /t >$ で囲み出力精度、その傾向を確認する。実験 8 では、プロンプト中に”Let’s think step by step.” と記述することで段階的に推論を行わせる Zero-shot CoT(Chain-of-Thought)[11] を行う。実験 1 から実験 6 の中で実験 3 の出力精度が最も低かったため、精度の向上を期待しデータセットは実験 3 のものを変更を加えず使用した。

3.6 実験結果

表 1 に実験結果を示す。実験結果として各実験での出力精度を示す。Fine-tuning により tuning 前のモデルより出力精度が向上したのは、実験 1 および実験 7 であり、その他の実験では Fine-tuning をした場合の方が精度は低下している。実験 1、実験 7 以外は Zero-shot learning による実験である。

表 1 各実験の出力精度

	tuning 前	tuning 後
実験 1	0.592 (71/120)	**0.708(85/120)**
実験 2	0.679 (38/56)	0.464 (26/56)
実験 3	0.679 (38/56)	0.357 (20/56)
実験 4	0.679 (38/56)	0.500 (28/56)
実験 5	0.679 (38/56)	0.607 (34/56)
実験 6	0.679 (38/56)	0.643 (36/56)
実験 7	0.600 (72/120)	**0.708 (85/120)**
実験 8	0.643(36/56)	0.357(20/56)

4 傾向分析

4.1 出力精度に対する考察

本節では、実際の実験結果から出力傾向を分析した結果とそれに伴う考察を示す。まず表 1 で示した出力精度について言及する。3.6 節で述べたように Fine-tuning を行うことで実験 1、実験 7 では精度の向上が見られ、それ以外の実験では精度の低下が見られた。これより、カタカナ語の意味推論タスクでは Fine-tuning は Few-shot learning で効果的だが、Zero-shot learning では精度向上に寄与しなかったといえる。

Zero-shot learning の精度低下は、訓練データの偏りと対象単語の文脈内使用例の希少性が原因と考えられる。前者は BCCWJ より抽出を行ったが、出現頻度を考慮せずに利用しており、多用される意味を

持つ対象単語が含まれる文章が訓練データに多く使用されている可能性が示唆されるためである。後者は BCCWJ 内で出現頻度が 1 である意味を持つ単語を含む文でテストデータを構成しているため、その語義を実際に文書等で使用する場面が少ないことが考えられる。例えば Zero-shot learning のテストデータ中の、”西洋料理で、順に出される一品。”という意味を持つコースという単語の正解選択肢の出力が全体的に見られなかった。食事に関する場面かつコース料理専門店等に限定されて使用されるため、使用したモデルのデータに含まれていない、もしくは少数しか実用例が含まれていない可能性がある。

実験 7 は実験 1 と比較して精度向上が見られず、タグ付与の有効性は確認できなかった。また実験 1 と実験 7 の判定結果のコサイン類似度は 0.7396 であり、出力の傾向が類似しているといえる。実験 8 でも実験 3 の精度向上を期待したが、Zero-shot CoT による精度向上に対する有効性は確認できなかった。これは訓練データとして利用したのが日本語データであることが原因である、もしくは連鎖的に回答を導くには、意味推論タスクでは効果が薄い可能性があるからであると推測している。また実験 3 と実験 8 の判定結果のコサイン類似度は 0.6949 であったため出力の傾向は類似しているといえる。

実験 4、実験 6 では実験 3、実験 5 と対応して比較すると、出力精度が向上していることがわかる。実験 4 は実験 3 の、実験 6 は実験 5 の訓練データを無作為に半分抽出したものを利用しており、データ数の増加がバイアスの増加を招いていると考えられる。さらに実験 2、3、5 を比較すると、出力精度は実験 3 が最低で実験 5 が最高であった。実験 3 はテストデータ中の対象単語を含む文のみから構成されており、実験 5 はテストデータ中の対象単語を含む文を実験 2 の訓練データから除いたものであるため、Zero-shot learning での意味推論では訓練データで対象単語の持つ意味を多数与えると精度が低下すると思われる。

4.2 出力傾向分析

出力の傾向を一部の例に絞って示す。まず実験 1、実験 7 について、Fine-tuning 前と後のどちらでも正解選択肢を出力しなかった例として、プラス、マイナスという単語が挙げられる。これらはデジタル大辞泉内で 9 個の意味区分が定義されており、詳細かつ限定的な語義に分割されている。この二つの単語に関しては正解選択肢の出力や精度向上はほとんど見られなかった。ただしプラスでは”加えること。加算。”、マイナスでは”よくないこと。また、悪い面。”という意味で Fine-Tuning により精度の向上が見られた。次に実験 2 から実験 6、実験 8(Zero-shot learning) に関して、正解選択肢よりも正しい出力を得られた例がある。ホームセンターの一部としてホームの正解を”家庭。家。「―バー」「マイ―」”とラベル付けしたが、Fine-tuning 前のモデルで”正確には「ホームセンター」という店舗の名称であるため、通常の「ホーム」とは異なる使われ方をしています。”というものが得られた。しかし Fine-tuning を行うことにより出力は正解選択肢となってしまった。

さらに実験 2 から実験 6 において”洋裁で、生地を裁つこと。裁断。：カット”, ”西洋料理で、順に出される一品。：コース”, ”建築・美術・音楽などの様式。型。：スタイル”, ”道具・機械などを組み立てて使えるようにすること。装置すること。：セット”, ”調べて、不都合なものが入り込むのを阻止すること。：チェック”, ”陰電気。また、その記号。：マイナス”, ”減じること。差し引くこと。：マイナス”, ”不利であること。不利益。損失。：マイナス”, ”ゴルフで、グリーン上に切られている目標の穴。：ホール”, ”企業組織のうち、局・部・課・係のような上下の組織。：ライン” の以上 10 個の語義が共通して、Fine-tuning 前後のモデルでどちらも正解選択肢を出力しなかった。どの語義も限定的に使用されるものであり、マイナスは特に精度は低かった (プラスは BCCWJ 内で出現頻度が 1 である語義は無いためテストデータに含まれていない)。

5 おわりに

カタカナ語の意味推論を行う複数の実験を行った結果、gpt-4o-mini-2024-07-18 モデルは限定的な場面で使用される意味を持つカタカナ語の意味推論は未だ精度が高くなく、さらに多数の意味を持つ単語では正しい語義を出力することが難しいことが確認できた。しかし Few-shot learning における Fine-tuning では出力の精度向上が見られたため、今回得られた出力精度の低い語義を含む文章を含めた適切な訓練データによって Fine-tuning を行ったモデルを利用すると、カタカナ語の意味推論におけるさらなる精度向上が得られる可能性が高い。

謝辞

本研究は JSPS 科研費 基盤研究 (C) 22K12161 の助成を受けたものです。

参考文献

[1] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language resources and evaluation**, Vol. 48, pp. 345–371, 2014.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, June 2019.

[3] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, **Proceedings of the 20th Chinese National Conference on Computational Linguistics**, pp. 1218–1227. Chinese Information Processing Society of China, August 2021.

[4] Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. Nibbling at the hard core of Word Sense Disambiguation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4724–4737, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[5] Terra Blevins and Luke Zettlemoyer. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1006–1017, Online, July 2020. Association for Computational Linguistics.

[6] Sakae Mizuki and Naoaki Okazaki. Semantic specialization for knowledge-based word sense disambiguation. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 3457–3470, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[7] Ming Wang and Yinglin Wang. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6229–6240, Online, November 2020. Association for Computational Linguistics.

[8] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szyd, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, BartKoptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Mi, Marcin Oleksy, Maciej Piasecki, Radliński, Konrad Wojtasik, StanisWoźniak, and PrzemysKazienko. Chatgpt: Jack of all trades, master of none. **Information Fusion**, Vol. 99, p. 101861, 2023.

[9] Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1562–1575, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[10] Edoardo Barba, Tommaso Pasini, and Roberto Navigli. Esc: Redesigning wsd with extractive sense comprehension. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4661–4672, 2021.

[11] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. **Advances in neural information processing systems**, Vol. 35, pp. 22199–22213, 2022.