カタカナ語の意味分類における Fine-Tuning の有効性検証：頻度と多義性が精度に与える影響

小滝主紀 [1] 佐々木稔 [1]
[1] 茨城大学 情報工学部
{24nm724g,minoru.sasaki.01}@vc.ibaraki.ac.jp

概要

LLM や PLM を用いたカタカナ語の意味分類では、和製英語や外来語といった独自の課題を持つ特性が精度に影響を与える可能性がある. そこで本研究では、BCCWJ から抽出したカタカナ語を対象として、DeBERTa V3 を Fine-Tuning し、精度向上への方策を探った. 対象単語を頻度と多義性で 4 象限に分類して分析した結果、Fine-Tuning によりベースラインから平均約 53%の精度向上を達成した. さらにベースラインモデルでは低頻度、低多義性語が、Fine-Tuning 後のモデルでは高頻度、低多義性語が最高精度であった. 統計分析により、頻度よりも多義性の方が精度に与える影響が大きいことが定量的に示された.

1 はじめに

大規模言語モデル (LLM) および事前学習言語モデル (PLM) に関する研究は、自然言語処理分野で活発に行われており、様々なタスクに適用する研究も進んでいる.WSD(語義曖昧性解消) もその一つである.WSD は文脈によって曖昧な対象単語の意味を正しく決定するタスクであり [1]、教師あり学習法と知識ベース法の 2 つの主要なアプローチがある. 教師あり学習法は、人間が対象となる多義語に正しい語義を付与したコーパスでモデルを訓練し、曖昧な単語に対して適切な語義を分類する. 知識ベース手法は、辞書やオントロジーなど外部知識を用いた分類手法である. これらの手法を基に LLM や生成 AI を WSD に使用する試みがなされている [2, 3, 4]. これらの研究は有望な性能を示しているが、まだ最先端レベルには達していない. ChatGPT を含む多くの LLM は主に英語データで訓練されており、日本語データは相対的に限られている. さらにカタカナ語には、英語を語源とする外来語や和製英語が含まれ、元来の意味と異なる場合があるため、文脈中の意味分類が正しく行われない可能性が高いという特有の課題を有する . これらの理由からカタカナ語の WSD は他の言語にはない特有の課題を有する. したがって本論文では日本語固有のデータで事前学習済みモデルを Fine-Tuning することにより、カタカナ語の意味分類精度を向上させ、その分類傾向を分析することを目的とする. 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[5] からデータセットを構築し、日本語データで追加学習した日本語 DeBERTa V3 モデル (ku-nlp/deberta-v3-base-japanese)[6, 7] をファインチューニングした. 対象単語の持つ頻度や多義性などの特性を分析し、カタカナ語 WSD の課題と改善の可能性を議論する. カタカナ語に焦点を当て、頻度と多義性の両方の観点から体系的に WSD 解消を試みた先行研究はほとんどない.

2 手法

ライセンス契約の条件により、本研究で用いた BCCWJ の全部または一部を再構成できるデータを研究成果で公表することは禁止されている. そのため、本論文では BCCWJ の内部構造を直接参照できる単語名や用例文などの情報を含まないことをあらかじめ明記しておく. 詳細な手順を以下に述べる.

2.1 カタカナ語の抽出

本研究におけるデータ抽出に用いたコーパスは書籍、雑誌、新聞、ウェブテキストなど多様なジャンルを網羅する日本語の均衡コーパスである BCCWJ を選定した. そして形態素情報付き XML データで、文の境界や形態素情報が構造化されている M-XML(Morphology-base XML) を、今回処理性の観点からデータセットの作成に採用した.BCCWJ では最小形態素単位を SUW(Short Unit Word) タグで提

供しているため、単語や語義のより詳細な傾向と統計分析を行えると判断し使用した.wType 属性が"外 (外来語)"の値を持つすべての文を抽出した結果、180,664 語、1,566,913 文が得られた. 次に WSD タスクに適さない語 (非標準語、固有名詞、感動詞、一つの用例文しかない単語) を除外するため、IPADIC を搭載した形態素解析器 MeCab[8] を使用した結果、最終的に 6,925 語が得られた. 最後に、実際の辞書に登録されている単語のみに焦点を当てるため、デジタル大辞泉 [9] を参照し、2 つ以上の語義定義を持つ単語とその語義を抽出した (1,639 語、801,899 文). これは、WSD タスクが本質的に複数の語義からの選択を必要とするためである.

2.2 アノテーション

先ほど抽出した 1,639 語を対象として訓練データのための語義予測データがアノテーションデータとして適切かどうかを検証するため、対象単語を含む 801,899 文からランダムに 200 文をサンプリングし、人手で正しい語義を付与し、5 つの GPT モデルで人間のアノテーションと予測結果を比較した. 以下の表 1 にその結果を示す.

表 1 人間と API による付与語義の一致率とコスト

モデル	一致率	出力コスト ($/1M)
gpt-4o	82.00%	10.00
gpt-4.1	81.50%	8.00
gpt-4.1-mini	**78.00%**	**1.60**
gpt-4o-mini	75.00%	0.60
gpt-4.1-nano	64.50%	0.40

gpt-4.1-mini は、最高精度であった gpt-4o と比較して 4%低い精度だがコストは約 1/6 であるため、78%という許容可能な精度を維持しながらコストを大幅に削減できると判断し選定した. Open AI API [1)] を使用した gpt-4.1-mini によって、1639 語を含む用例文の語義を予測させた際にプログラムで実際に使用したプロンプト形式を図 1 に示す.

```
prompt = (
    f"For each of the following sentences, select only the number of the
    option that most closely matches the meaning of [target word].\n"
    f"[Options]\nContext: [Example Sentence]"
)
messages = [
    {"role": "system", "content": "You are an AI that outputs only the
    correct answer number based on the options."},
    {"role": "user", "content": prompt}
]
```

図 1 語義予測のプロンプトテンプレート

図 1 に示される選択肢は各対象の語義を表しており、各選択肢には番号が割り当てられているため、モデルはこの番号の中から予測するように調整した. 得られた予測結果はアノテーションデータとして使用し、その後の Fine-Tuning 用データの作成に活用した.

2.3 Fine-Tuning

2.3.1 データ特徴分析

Fine-Tuning を行う前に、先ほど得たアノテーションデータの内容を調査した. まず、tiktoken トークナイザーを使用して各単語をトークン化し、サブワードに分割されなかった単語を抽出した (940 語、504,688 文). 抽出された単語は、単語通りの形で理解されているため、十分な文脈パターンを学習していると解釈できる. この 940 語について BCCWJ に用例文がいくつ存在するかを示す頻度を計算した. 分布の平均は 536.90、中央値は 169.00、標準偏差は 1,168.16 であった. 平均が中央値より大きく、分布が右に大きく歪んでいることから、ジップの法則 [10] に従っている可能性が示唆され、実際に頻度上位 24.5%の 230 語が全文の 80.0%（403,819/504,688）を占めていることが判明した. したがって、データはジップの法則に従い、パレートの法則を適用できると判断し、これを Fine-Tuning のために使用した. さらにここで得られた単語の頻度とエントロピーを計算した. ここでエントロピーとは情報の不確実性を表す尺度を示す. 語義を単語に含まれる情報と定義すると、語義の曖昧さは単語のエントロピーによって定量化できる. エントロピーが高いほど、語義分布がより均等に分布し予測が困難であることを示し、単語の多義性が大きいことを表す. シャノンのエントロピー [11]

$$H = -\sum_{i} p(i) \log_2 p(i). \tag{1}$$

を計算して各語のエントロピー値とした.230 語の分布における頻度の中央値は 1,057.50、エントロピーの中央値は 0.8968 であった. 得られた中央値を閾値として 230 語を頻度と多義性の 2 軸で分類した結果、以下の表 2 に示すように 4 つの象限に分割することができた.

表 2 頻度と多義性による分割象限

	高多義性	**低多義性**
高頻度	[Q1] 61 words (26.5%)	[Q2] 54 words (23.5%)
低頻度	[Q3] 54 words (23.5%)	[Q4] 61 words (26.5%)

Q3(高多義性・低頻度) が最低精度、Q2(低多義

1） https://platform.openai.com/docs/

性・高頻度) が最高精度になると予測した. これは高多義性により語義の曖昧性が増加し、高頻度であることがその単語が多くの文脈で使用され、一般的に重要な単語であると予測できたためである. この仮説を検証するため、各象限について DeBERTa V3(ku-nlp/deberta-v3-base-japanese) モデルを Fine-Tuning し、その精度を測定する.

2.3.2 訓練手順

各象限において、各単語の比率を維持したまま訓練データを成形するため、層化分割によって訓練、評価、テストデータを 8:1:1 の比率でランダムに分割した. ベースラインのモデルとファインチューニングモデルの精度を平均値で比較するため、ランダムシード値 42-46(再現性確保) で 5 回分割を実施した. この過程で、アノテーション中に有効なラベルが与えられなかったため、一部の例文がスキップされた. しかし、スキップ率は全体の 0.096%(148/153,713) に過ぎないため、データ分析と精度への影響は小さいとして使用しなかった. こうして作成した訓練、開発、テストデータを、訓練中に容易に読み込める形式に整形し、プロンプト形式の JSONL ファイルを生成した (図 2).

```
{
  "text": "Please determine the meaning of the word [target word].\n\
    nChoices:\n0: Meaning 1\n1: Meaning 2\n2: Meaning 3\n\nsentences:
    Example Sentences",
  "label": Correct Meaning Label,
  "word": "target word",
  "original_sentence": "Example Sentences",
  "label_text": "Correct Meaning"
}
```

図 2 Fine-Tuning 時のプロンプトテンプレート

Transformers 4.30.2 と PyTorch2.0.1 を使用して ku-nlp/deberta-v3-base-japanese モデルを Fine-Tuning した.5 つの異なるランダムシード (42-46) で各象限において Fine-Tuning モデルを 5 つ作成し、統計的信頼性を確保した.

3 結果

上記の手順に従ってファインチューニングを実施し、各象限でモデル性能を評価した. 表 3 はベースラインモデルの性能、表 4 は Fine-Tuning モデルの性能を示す. 性能の評価指標は Accuracy、Precision、Recall、F1 スコアとした. 各値は 5 回計測した平均値である.

表 3 DeBERTa V3 における性能

	Acc.	Pre.	Rec.	F1
Q1	0.2806	0.3157	0.2806	0.2881
Q2	0.3326	0.3723	0.3326	0.3470
Q3	0.2901	0.2962	0.2901	0.2895
Q4	**0.4624**	**0.4896**	**0.4624**	**0.4666**

表 4 Fine-Tuning モデルにおける性能

	Acc.	Pre.	Rec.	F1
Q1	0.8343	0.8334	0.8343	0.8336
Q2	**0.9393**	**0.9394**	**0.9393**	**0.9393**
Q3	0.8114	0.8206	0.8114	0.8128
Q4	0.9158	0.9162	0.9158	0.9159

表 3 に示すように、ベースラインモデルでは低頻度・低多義性の Q4 が約 46%の最高精度を達成した. 低多義性象限 Q2、Q4 は、高多義性象限 Q1、Q3 より数%高い値を示した. この結果は、多義性が低いほどベースラインモデルでは高い精度を達成することを示している. 一方、表 4 に示すように、高頻度・低多義性の Q2 が約 93%の最高精度を達成した. 逆に、低頻度・高多義性の Q3 が最低精度を記録した. これらの結果は、2.3.1 節で述べた仮説と一致した. さらに、低多義性象限 Q2、Q4 は高多義性象限 Q1、Q3 より約 8-12%高い精度を示し、多義性が性能に与える影響を明確に示している. これらの差の統計的有意性を評価するため、10,000 回のリサンプリングによるブートストラップ分析 [12] を実施した. 以下の表 5 に結果を示す.

表 5 各象限間の統計的差と信頼区間.

ある特性下の比較	平均値の差	95%信頼区間
Q1-Q2 (高頻度下の多義性)	-10.50%	[-10.83, -10.19]
Q3-Q4 (低頻度下の多義性)	-10.44%	[-11.41, -9.51]
Q1-Q3 (高多義性下の頻度)	+2.29%	[+1.40, +3.20]
Q2-Q4 (低多義性下の頻度)	+2.35%	[+1.88, +2.74]

95%信頼区間に 0 が含まれていないため、すべての象限間で統計的に有意な差が示された.

4 傾向分析

結果が示すように、ベースラインモデルは低頻度・低多義性の Q4 で最高精度を達成し、Fine-Tuning モデルは高頻度・低多義性の Q2 で最高精度を達成した. これらについてまず Fine-Tuning モデルについて考察を述べる. 頻度に関しては、2.3.1 節で述べたように、多くの文脈で使用される重要な単語であることに起因すると考えられる. これは訓練中に頻出し、モデルがその意味表現を十分に獲得できることを意味する. そのため高頻度であることが精度向上

へ寄与したと考えられる. 多義性に関しては、多義性が高いほど曖昧さが増すと予測できる. 実際に高多義性象限 Q1、Q3 は低多義性象限 Q2、Q4 より精度が低く、予測が曖昧であることが確認された. これは対象単語が持つ語義が多いほど、各語義に関連する文脈パターンが断片化され、文脈に沿った語義の対応関係が不明瞭になることに起因すると考えられる. したがって、Q1 が高頻度であるにもかかわらず Q2 より大幅に低い精度を示した理由は、その高多義性にあり、高頻度単語は意味分類における精度向上に寄与するが、低多義性はさらに重要であり、精度向上により大きく貢献することを意味する.3 節で示された象限間の統計的有意な差は、この多義性が精度に与える影響は頻度より大きい事実を定量的に実証している. 次にベースラインモデルについて考察を述べる.Fine-Tuning モデルとは異なり、ベースラインモデルは Q2 ではなく Q4 で最高精度を達成した. 高頻度であることは多様な文脈で使用されるため、未訓練状態では予測が分散する可能性がある一方で、低頻度語は訓練前のサンプル数が少なく、特定の文脈でのみ使用され、文脈パターンが明確になる可能性がある. 実際、低多義性の Q2 と Q4 を比較すると、低頻度の Q4 がより高い精度を示した. 同様に、高多義性の Q1 と Q3 を比較すると、低頻度の Q3 がより高い精度を示した. これは、低頻度であることがより明確な文脈パターンを持ち、精度を向上させるという仮説を支持している. このように頻度が精度に与える影響はベースラインモデルと Fine-Tuning モデルで逆転する. これは、Fine-Tuning により高頻度単語の豊富な訓練サンプルを活用でき、文脈の多様性が有利に働くことを示唆している. これは、カタカナ WSD タスクにおける高頻度単語の正確な意味分類には、Fine-Tuning などの追加学習が非常に重要であることを示している. 次に評価指標について考察する. ベースラインモデルでは、Precision が Accuracy と Recall より高かった. これは、モデルが予測しやすい特定の語義に偏って予測する傾向があり、予測分布に偏りがあることを示唆している.Fine-Tuning モデルでは、4 つの指標がほぼ一致しており、モデルがほぼ均等に語義を予測できていることを示している. ベースラインモデルと比較して、全体で平均約 45%から 61%の大幅な精度向上が確認された. これは本研究で使用したデータセットが高バランス、高品質であり、ベースラインモデルの予測バランスの改善に貢献したことを示している. 不均衡な訓練例が特に非英語言語で誤った語義予測を引き起こす可能性があることが先行研究で示されており [13]、本研究のデータバランス改善が精度向上に大きく寄与したという知見は補強される. さらに Fine-Tuning が多義語の語義分離にどう影響するかを確認するために、230 単語について Fine-Tuning 時に使用したテストデータを用いて各用例文の埋め込みベクトル (平均プーリング) を計算した. 付与されている語義ラベルに基づき、同一語義ラベルを持つ文を同語義ペア (4,824,903 ペア)、異なる語義ラベルを持つ文を異語義ペア (2,793,458 ペア) として区分し、ベースラインモデルと各象限の Fine-Tuning モデルでコサイン類似度を比較した. その結果、Fine-Tuning 後は同語義ペアで平均 0.045、異語義ペアで平均 0.092 の低下が見られた. 特に Q2 の異語義ペアでは 0.144 と最大の低下を示した. また、ベースラインモデルでは同語義ペアと異語義ペアの差がわずか 0.002〜0.007 であったのに対し、Fine-Tuning モデルでは最大 0.096(Q2) まで拡大した. この結果は、Fine-Tuning により異なる語義を持つ用例文がベクトル空間上でより分離されるようになったことを示しており、モデルがカタカナ語の細かな語義ニュアンスの違いをより適切に捉えられるようになったといえる.

5 おわりに

本研究では BCCWJ から抽出したカタカナ語 230 語と、OpenAI API を使用してアノテーションされたこれらの対象語を含む 403,819 文を使用して、DeBERTa V3 を Fine-Tuning し、カタカナ語の意味分類タスクの性能を評価した. 評価結果から、バランスの取れたデータを用いた際の Fine-Tuning が精度向上に対し有効であること、多義性の方が頻度よりも精度向上への影響が非常に大きいこと、ベースラインモデルと Fine-Tuning モデルにおいて頻度の影響が逆転し、文脈の多様性が語義曖昧性に対し有効に働くようになることが考察できた. 本研究の知見は、カタカナ語の意味分類タスクにおける Fine-Tuning の重要性を実証し、また頻度と多義性が精度に与える影響を定量的に明らかにし、多義性が頻度よりも精度向上に大きく寄与することを示した. 今後の課題としては、人手アノテーションによる信頼性を担保した検証、頻度と多義性以外の精度向上に寄与する要因の発見と分析、より広範なデータを使用したデータセットの作成がある.

謝辞

本研究は JSPS 科研費 22K12161, 25K15242 の助成を受けたものです.

参考文献

[1] Roberto Navigli. Word sense disambiguation: A survey. **ACM Computing Surveys**, Vol. 41, No. 2, p. Article 10, 2009. https://doi.org/10.1145/1459352.1459355.

[2] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. Chatgpt: Jack of all trades, master of none. **Information Fusion**, Vol. 99, p. 101861, 2023. https://doi.org/10.1016/j.inffus.2023.101861.

[3] Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 1562–1575, 2024. https://doi.org/10.18653/v1/2024.eacl-long.94.

[4] Van-Hien Tran, Raj Dabre, Heang Kaing, Haiyue Song, Hideki Tanaka, and Masao Utiyama. Exploiting word sense disambiguation in large language models for machine translation. In **Proceedings of the First Workshop on Language Models for Low-Resource Languages**, pp. 135–144, 2025. https://aclanthology.org/2025.loreslm-1.10/.

[5] 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. https://doi.org/10.1007/s10579-013-9261-0.

[6] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In **Proceedings of the 9th International Conference on Learning Representations (ICLR)**, 2021.

[7] Kyoto University NLP Group. Deberta v3 japanese base model, 2023. https://huggingface.co/ku-nlp/deberta-v3-base-japanese.

[8] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**, pp. 230–237, 2004.

[9] Shogakukan Dictionary Editorial Department. Digital daijisen, 2012. https://www.weblio.jp/.

[10] George Kingsley Zipf. **Human Behavior and the Principle of Least Effort**. Addison-Wesley, 1949.

[11] Claude E. Shannon. A mathematical theory of communication. **Bell System Technical Journal**, Vol. 27, No. 3, pp. 379–423, 1948.

[12] Bradley Efron and Robert J. Tibshirani. **An Introduction to the Bootstrap**. Chapman & Hall/CRC, 1993.

[13] Dilan Sumanathilaka, Nicholas Micallef, and Julian Hough. Prompt balance matters: Understanding how imbalanced few-shot learning affects multilingual sense disambiguation in llms, 2025. arXiv preprint arXiv:2510.03762.