大規模言語モデルによる固有表現理解の粒度に関する検討

前木場 舜 [1]　土屋 雅稔 [1]
[1] 豊橋技術科学大学大学院 情報・知能工学専攻
{maekoba, tsuchiya}@is.cs.tut.ac.jp

概要

人間は概念を理解する際，対象を粗い分類から細かい分類へと段階的に捉える階層的な意味構造を自然に用いている．固有表現も，人名・地名・製品名など多様な種類を持ち，意味の粒度を持つ階層構造として体系化されている．一方，大規模言語モデル（Large Language Model; LLM）は，このような人間が設計した概念体系を明示的に学習しているわけではなく，固有表現をどの粒度で概念的に理解しているかは明らかではない．本研究では，拡張固有表現（Extended Named Entities; ENE）の階層構造に注目し，LLM による固有表現理解の粒度を検討する．

1　はじめに

近年，大規模言語モデル（Large Language Model; LLM）は質問応答や文章生成など多様な自然言語処理タスクにおいて高い性能を示しており，人間に近い言語運用能力を獲得しつつあると評価されている [1, 2]．一方で，LLM は人間のように概念体系や意味構造を明示的に学習するのではなく，大量のテキストデータから統計的な関連性を通じて言語表現を獲得している．このため，LLM が言語表現をどのような単位で捉え，どの程度概念的に理解しているのかは必ずしも明らかではない．

人間の言語理解では，対象を粗い分類から細かな分類へと整理する階層的な概念構造が重要な役割を果たす．固有表現もその一例であり，人名，地名，製品名といった分類は，意味の粒度を持つ階層構造として整理することができる．拡張固有表現（Extended Named Entities; ENE）は，このような人間の概念的整理を反映した体系として，固有表現を階層的に分類する枠組みを提供している [3]．

しかし，LLM はこのような階層的な概念体系を明示的に与えられて学習しているわけではなく，固有表現についても，それらがどの粒度の概念として捉えられているのかは十分に検証されていない．すなわち，LLM が固有表現を表層的な語として扱っているのか，あるいは概念的な位置づけまで含めて理解しているのかは不明である．

入力：
文脈（context）
私は愛知県の豊橋市に住んでいます．
質問（question）
文中に含まれる ○○ を答えてください．

質問形式：

名前のみ	○○：地名
定義のみ	○○：場所に関する名前．
名前＋定義	○○：地名（場所に関する名前．）

出力：
抽出された固有表現：愛知県，豊橋市

図 1　入力（文脈＋質問）と出力の関係

本研究では，人間が設計した概念体系である ENE の階層構造を LLM の固有表現理解を観察するための評価軸として用いる．ENE は意味の粒度を段階的に表現しており，この構造に基づいて LLM の応答を分析することで，固有表現がどの粒度まで概念的に扱われているのかを捉えることができると考えられる．本研究では，「理解」を ENE 階層上で一貫したカテゴリ選択が行われるという操作的な振る舞いとして定義し，LLM の固有表現理解のあり方を探索的に観察することを目的とする．

2　関連研究

LLM の能力分析に関する研究は近年急速に進展しており，概念理解や分類的推論の性質に着目した分析も数多く報告されている．本節では，LLM の概念理解，固有表現認識，および階層的固有表現体系に関する研究を概観し，語彙レベルの分析が中心である点で本研究と異なる位置づけを整理する．

LLM は，明示的な概念体系や分類規則を与えられていないにもかかわらず，語彙間の上位・下位関係や包含関係を一定程度正しく扱えることが報告されている [4, 5]．これらの研究では，上位概念への

分類は比較的安定して行える一方で，下位概念の細かな区別は難しい傾向があることが示されており，LLM の概念的振る舞いが扱う概念の粒度と関係している可能性が示唆される．一方で，固有表現のような言語的対象を，概念の粒度や階層構造の観点から体系的に分析した研究は限られている [6]．

LLM を固有表現認識（NER）に適用する研究では，ChatGPT や GPT-4 がゼロショット条件下でも基本的な固有表現を抽出可能であることが示されている [7, 8]．一方で，専用に学習された NER モデルと比較すると，厳密一致に基づく評価では性能が劣る場合があることも報告されている [7]．また，LLM の出力に対する誤り分析では，過剰抽出，検出漏れ，カテゴリの取り違えなどの誤りが指摘されており，特に固有表現の境界やカテゴリが曖昧な場合に誤りが生じやすいことが示されている [7, 8]．

ENE は，固有表現を意味的観点から整理し，階層構造として定義した体系である [3]．ENE は，人名や地名といった従来の大分類に加え，より細かな意味的カテゴリを段階的に区別できる点に特徴がある．また，日本語における固有表現アノテーションの枠組みとして，現代日本語書き言葉均衡コーパス（BCCWJ）をはじめとするコーパス構築にも適用されている [9]．

3 データセットとタスク設定

3.1 データセット

本研究では，日本語の大規模コーパスである現代日本語書き言葉均衡コーパス（BCCWJ）を用いる [9]．BCCWJ は書籍，新聞，雑誌，ウェブテキストなど，多様なジャンルから収集された日本語テキストを含んでおり，現代日本語の使用実態を広く反映している点に特徴がある．

本研究では，BCCWJ に付与された ENE アノテーションを利用する．ENE は固有表現を階層的に整理した体系であり，人名や地名といった粗い分類から，より細かな意味的カテゴリまでを包含している．このような階層構造を持つ固有表現体系を備えた日本語データセットは限られており，LLM の固有表現理解を粒度の観点から分析するための基盤として適している．

本研究では，ENE の第 2 階層から第 4 階層までを対象とし，第 4 階層の固有表現出現を起点として QA データセットを構築した．各固有表現出現に対して，対応する上位階層（第 2〜第 4 階層）のカテゴリを遡って取得し，同一の文脈から階層ごとの質問を生成している．その結果，すべての階層は共通の 17,456 文の文脈集合に基づいて構成されている．また，ENE の階層構造および各階層におけるカテゴリ種類数を表 1 に示す．

表 1 ENE 階層別のカテゴリ種類数

階層	カテゴリ種類数
第 1 階層	1
第 2 階層	11
第 3 階層	57
第 4 階層	120

3.2 タスク設定

本研究では，入力文に含まれる固有表現を抽出し，それぞれの固有表現カテゴリを予測させるタスクを設定する．ただし，本研究の目的は，固有表現認識性能の向上や既存手法との精度比較ではない．

入力文には，1 文中に複数の固有表現が含まれる場合も多く，LLM はそれらを同時に予測する必要がある．この設定では，すべての固有表現を正確に予測できなくとも，一部の固有表現について正しい判断が行われる場合がある．そのため，本研究では，LLM の出力を単一の正解ラベルに還元するのではなく，文ごとに予測された固有表現の集合として扱う．また，ENE の階層構造を考慮し，固有表現カテゴリを複数の階層レベルに分けて評価・分析することで，LLM が固有表現をどの程度の粒度まで区別できているのかを段階的に観察する．

3.2.1 入力条件とプロンプト構成

本研究では，LLM に与える入力情報の違いが固有表現カテゴリ予測に与える影響を分析するため，入力条件として「名前のみ」「定義のみ」「名前＋定義」の 3 条件を設定する．いずれの条件においても，モデルに与える指示文，出力形式，および候補となる ENE カテゴリ集合は共通とし，入力として提示するカテゴリ情報のみを切り替えている．

図 1 に，各入力条件におけるプロンプト構成の概要を示す．本図では，文脈と質問からなる入力の構造と，条件ごとに付与されるカテゴリ情報の違いが把握できるよう，プロンプトの入力部分および対応する出力例を抜粋して示している．また，実験で用いたプロンプトの全文は，付録 A.2 の通りである．

なお，本研究で用いる「定義」は，ENE において各カテゴリに付与されているカテゴリ定義文に基づいている [3]．定義のみ条件および名前＋定義条件では，対応する ENE カテゴリ名に紐づく定義文をそのままプロンプトに付与し，内容の改変や追加は行っていない．

4　評価方法

本研究では，LLM が出力する固有表現の理解を評価するため，予測結果と正解アノテーションの対応関係に基づく評価を行う．1 文中に複数の固有表現が含まれる状況を考慮し，単一ラベルによる評価ではなく，文ごとに得られる固有表現の集合を評価単位とする．

4.1　評価指標

各入力文に対して，LLM が出力した固有表現とカテゴリの組を予測集合，BCCWJ に付与された ENE アノテーションに基づく固有表現とカテゴリの組を正解集合と定義する．予測集合と正解集合の対応関係に基づき，真陽性（TP），偽陽性（FP），偽陰性（FN）を集計する．

評価指標として，これらに基づく precision，recall，および F1 スコアを用いる．完全一致に基づく accuracy は，一部のみ正しい予測が行われた場合の挙動を適切に反映できないため，本研究では採用しない．

4.2　階層レベル別評価

ENE の階層構造を考慮し，第 2〜第 4 階層それぞれにおいて TP，FP，FN を集計し，階層レベル別に F1 スコアを算出する．これにより，固有表現カテゴリの粒度の違いに応じた LLM の予測傾向を分析する．

5　実験結果

本節では，入力条件（名前のみ，定義のみ，名前＋定義）および ENE の階層レベル（第 2〜第 4 階層）の違いが，固有表現カテゴリ予測の F1 スコアに与える影響を分析する．以降の平均値は，カテゴリ間の粒度差に着目するため，出現頻度の影響を抑える目的で第 4 階層カテゴリ（120 カテゴリ）を単位としたマクロ平均として算出した．

表 2　階層レベル別の平均 F1 スコア（%）

階層	名前	定義	名前＋定義
第 2 階層	51.37	54.56	52.86
第 3 階層	64.63	65.83	67.06
第 4 階層	69.73	68.63	72.30

5.1　階層レベル別の全体傾向

表 2 に，階層レベル別の平均 F1 スコアを示す．第 2 階層では定義のみ条件が最も高い F1（54.56）を示した一方，第 3 階層および第 4 階層では名前＋定義条件が最も高い F1（第 3 階層：67.06，第 4 階層：72.30）を示した．この結果から，入力として与える情報（名前・定義）の有効性は，予測対象とするカテゴリの粒度（階層レベル）に応じて異なる可能性が示唆される．

5.2　条件間比較によるカテゴリ差異

第 4 階層カテゴリごとに条件間の優劣を比較した結果，名前のみ条件が定義のみ条件を上回るカテゴリは 46，定義のみ条件が名前のみ条件を上回るカテゴリは 43，同率のカテゴリは 31 であり，一様な優劣は見られなかった．また，名前＋定義条件は多くのカテゴリで名前のみ条件を上回る一方，名前のみ条件を下回るカテゴリも 37 存在した．このことから，定義情報の付与は必ずしも一貫して性能向上をもたらすわけではなく，カテゴリに応じて有効な手がかりが異なる可能性が示される．

5.3　条件差が顕著なカテゴリ例

条件差が大きいカテゴリの具体例を示すため，第 4 階層において条件間の F1 スコア差が 10 ポイント以上のカテゴリを抽出し，設問数 30 以上のものに限定して例示する（表 3）．

表 3 は，(1) 名前のみ条件が定義のみ条件を上回る例，(2) 定義のみ条件が名前のみ条件を上回る例，(3) 名前のみ条件が名前＋定義条件を上回る例をそれぞれ示したものである．本表は，条件差が生じうる具体例を示すことを目的としており，各カテゴリにおける傾向を一般化するものではない．

なお，定義のみ条件と名前＋定義条件の比較については，全体的な傾向が前節の集計結果に示されている一方，本節では，定義情報の付与が必ずしも性能向上に寄与しないという直観に反する挙動を明確に示すため，「名前のみ」と「名前＋定義」の比較に

表 3　第 4 階層で条件間 F1 差が 10 ポイント以上のカテゴリ例（%）

例	第 4 階層	n	比較元	比較先
名前 > 定義	地位職業名	3287	49.15	34.99
名前 > 定義	政党名	162	77.49	63.06
名前 > 定義	都道府県州名	476	64.73	53.27
定義 > 名前	国籍名	146	37.07	15.79
定義 > 名前	法人名_その他	182	73.63	59.73
定義 > 名前	事故事件名_その他	91	72.46	56.74
名前 > 名前＋定義	戦争名	55	94.74	86.89
名前 > 名前＋定義	会議名	81	73.80	68.89
名前 > 名前＋定義	条約名	40	87.06	82.35

焦点を当てた．

6　考察

本節では，前節の実験結果（表 2〜表 3）を踏まえ，入力条件（名前・定義）の違いが LLM の固有表現カテゴリ予測に与える影響と，ENE 階層構造との関係について考察する．

6.1　入力情報と粒度の関係

表 2 より，第 2 階層では定義のみ条件が最も高い一方，第 3・第 4 階層では名前＋定義条件が最も高い．この結果は，固有表現カテゴリ予測において，定義が広い概念の同定（粗い粒度）を助ける場合がある一方で，細かな粒度では名前と定義の併用が有効となる可能性を示唆する．なお，本研究の平均値は第 4 階層カテゴリ単位のマクロ平均であり，ここでは階層間の厳密比較ではなく，傾向の観察にとどめる．この傾向は，LLM が上位概念は比較的安定して扱える一方で，下位概念の区別では誤りが生じやすいという既存の報告 [5, 4] とも整合的である．

6.2　カテゴリによる有効情報の違い

前節の条件間比較結果より，第 4 階層では名前のみが定義のみを上回るカテゴリと，定義のみが名前のみを上回るカテゴリの双方が多数存在し，一様な優劣は見られなかった．このことは，LLM がカテゴリごとに異なる手がかりを用いて予測を行っている可能性を示唆する．

例えば，名前が有利な例（表 3）では，名称自体にカテゴリを示唆する語彙的手がかりが含まれている，あるいは固有名詞の頻度や典型性が高いことにより，名前のみでもラベル同定が可能である状況が考えられる．一方，定義が有利な例では，「〜その他」のように包含範囲が広いカテゴリや，名前のみから境界が推測しにくいカテゴリにおいて，定義文が補助的特徴として機能した可能性がある．

6.3　定義付与による性能低下の要因

表 3 に示すように，名前のみが名前＋定義を上回るカテゴリも観察され，定義の付与が常に有益とは限らない．表 3 に示した例からは，少なくとも以下の可能性が考えられる．

第一に，定義文が一般的・上位概念的な語彙を含む場合，モデルがより粗い概念へ引き寄せられ，本来の細分類ラベルの選択が弱まる可能性がある．第二に，定義文が複数カテゴリに共通する特徴を含む場合，判別情報としてはノイズとなり，名前に含まれる強い手がかりを相対的に減衰させる可能性がある．第三に，定義文の記述が短く抽象的である場合，モデル内部の既存知識と整合しない解釈が生じ，結果としてラベル選択が不安定になる可能性もある [6]．これらはいずれも観察に基づく仮説であり，定義文中の語彙特徴や誤りタイプ（上位カテゴリへの丸め，近接カテゴリへの混同など）を体系的に分析することで検証が可能である．

7　おわりに

本研究では，ENE の階層構造を評価軸として導入し，粒度の異なる固有表現カテゴリにおける F1 を比較した．その結果，入力条件やカテゴリによって最も高い性能を示す階層レベルが異なり，LLM の振る舞いは単一の粒度では捉えにくいことが示唆された．これは，固有表現が粒度を持つ概念体系であるという観点から，LLM がどの粒度で判断しているかを観察する枠組みとして，ENE が有用である可能性を示すものである．一方で，階層間の比較は集計単位やラベル設計の影響も受けうるため，同一設問集合に対する階層間の誤り遷移を追跡するなど，粒度の観点から誤り構造をより直接に分析する必要がある．

本研究にはいくつかの限界がある．対象モデルが単一であること，プロンプト条件が限定的であること，および集合一致に基づく F1 により誤りの質を直接区別していない点である．今後は，定義文の形式的差異や ENE 階層上の距離に基づく誤り分析，ならびに複数モデル・複数設定での再現確認を通じて，LLM が固有表現をどの粒度で概念化しているかをより詳細に検証することが課題である．

謝辞

本研究の一部は，JSPS 科研費 JP22K12167 の助成を受けたものです．

参考文献

[1] Tom B. Brown, et al. Language models are few-shot learners. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2020.

[2] OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023.

[3] Satoshi Sekine, et al. Extended named entity ontology with attributes. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)**, 2008.

[4] Qiang Ji, et al. LLM-Hype: A targeted evaluation framework for hypernym-hyponym identification in large language models. In **Proceedings of the VLDB Workshops**, 2025.

[5] Vladimir Moskvoretskii, et al. Are LLMs good at lexical semantics? a case of taxonomy learning. In **Proceedings of LREC-COLING**, 2024.

[6] Fabio Petroni, et al. Language models as knowledge bases? In **Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, 2019.

[7] Rui Han, et al. Is information extraction solved by ChatGPT? an analysis of performance, evaluation criteria, robustness and errors. **arXiv preprint arXiv:2305.14450**, 2023.

[8] Zhengnan Xie, et al. Decomposed prompting for named entity recognition with large language models. In **Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2023.

[9] Kikuo Maekawa, et al. Balanced corpus of contemporary written japanese. **Language Resources and Evaluation**, 2014.

A　付録

A.1　実験設定の詳細

本研究では，以下の LLM を用いて実験を行った．

- llm-jp/llm-jp-3.1-13b-instruct4

推論時の主な設定は以下の通りである．

- max_new_tokens: 200
- num_beams: 5
- do_sample: False

各文は 1 文ずつモデルに入力し，バッチ処理は行っていない．すべて zero-shot 設定で実行した．

A.2　プロンプトの構造と全文

本研究で用いたプロンプトは，「文脈」「質問」「出力形式指定」から構成される．以下に，本研究で使用したプロンプト全文を示す．

あなたのタスクは、以下の文脈に含まれる固有表現を抽出することです。
質問で求められた対象の固有表現を、文脈からそのまま抜き出してください。
答えはできるだけ短く、単語のみを記載してください。
複数ある場合は、カンマ (,) で区切ってください。
答えのみを記載し、補足や説明は一切しないでください。
マークダウン形式は使わないでください。
「抽出された固有表現:」の後に、必ず回答を記載してください。

文脈: `{sample['context']}`
質問: `{sample['question']}`
抽出された固有表現:

入力条件（名前のみ・定義のみ・名前＋定義）の違いは，質問文中にカテゴリ名またはその定義文を付与するか否かのみであり，指示文および出力形式はすべての条件で共通である．

A.3　データ前処理

文分割には `ja_sentence_segmenter` を用い，正規化（neologd），改行分割，句読点分割を順に適用した．1 文中に複数の固有表現が含まれる場合，正解集合および予測集合はいずれも集合として扱い，評価時に一致判定を行った．

A.4　評価方法の具体例

例文「私は愛知県の豊橋市に住んでいます.」に対し，地名抽出を行う場合を考える．正解集合が {愛知県, 豊橋市}，予測集合が {愛知県, 豊橋市} のとき，TP は {愛知県, 豊橋市}，FP および FN は空集合となる．