大規模言語モデルにおける語彙関数知識の類推推論による検討

楊 宇軒 [1] 郭 凱 [1] 河野 優輝 [2] ルパージュ イヴ [1]
[1] 早稲田大学 大学院情報生産システム研究科 [2] 旭川工業高等専門学校
yang98@asagi.waseda.jp, guokai@akane.waseda.jp,
p248007@edu.asahikawa-nct.ac.jp, yves.lepage@waseda.jp

概要

本研究の目的は、意味・テキスト理論で定義されている語彙関数が、大規模言語モデルにおいてどの程度習得されているかを評価することである。複数のオープンソース大規模言語モデルを対象に、特定の語彙関数に関連するデータセットを用いたタスクを実行し、それらの正確率を統計的に分析した。実験結果から、大規模言語モデルは異なる語彙関数に対して正確率に明確な差があることを示した。このことは、モデルが異なる意味関係を理解する能力に違いがあることを示唆している。また、大規模言語モデルのサイズと正確率には強い相関関係が見られることも明らかになった。

1 はじめに

機械学習と自然言語処理の分野において、類推タスクはモデルの言語理解能力および言語生成能力を評価するための重要な手法の一つと見なされてきた。類推関係の課題では、モデルが適切な解を推論することで、語彙内に存在する単語関係に対するモデル理解度を検討できる。有名な例として、「王」と「妃」の関係が「男性」と「女性」の関係に等しいということに基づき、モデルに「男性：女性：：王：x」のような課題を与えた際、「妃」を出力できれば、モデルが複雑な言語現象を適切に処理できていると結論付けられる。 そして、モデルが語彙関数 **Son** や **Anti** を処理する際に異なる挙動を示すことや、多言語環境での処理に違いが見られることが挙げられる。類推タスクは、語彙の理解において重要な役割を果たすだけでなく、モデルの汎化能力や言語横断的な転移学習能力の評価にも寄与しており、より高度で適応性のある自然言語処理システムの構築においても重要な役割を担っている。本論文は、オープンソースの大規模言語モデルが類推課題において、どのような性能を示すかを探求することを目的とし、意味・テキスト理論 [1] の枠組みを用いて評価する。具体的には、異なる語彙関数に対して多言語での類推タスクを実施し、それらの正確性や相関性を比較する。本研究を通じて、異なる言語モデルや語彙関数が類推結果に及ぼす影響を説明し、今後のモデル設計と最適化への支援をを提供することを目指す。また、モデルに中国語と日本語のデータセットを処理させ、異なる言語間における汎化能力を評価する。

2 実験データと実験方法

意味・テキスト理論に基づく語彙関数の概念を用いてデータセットを作成し、それを用いて実験を行った。

2.1 意味・テキスト理論の紹介

語彙関数（仏：fonction lexicale）は、意味・テキスト理論（仏：théorie sens-texte (TST)）の枠組みで開発された概念である [2]。これは、特定の語彙単位（仏：unité lexicale (UL)）間の意味関係、特に共起語（コロケーション）や語彙派生を記述し、体系化するために使用される [3]。また、語彙関数は技術用語（説明的な組合せ辞典）の構築にも用いられ、ある種の統語的表現における抽象的なノードとして機能する。基本的に、語彙関数は $f()$ という形式をとり 、語彙表現の集合 $f(L)$ と語彙単位 L との対応関係を表す。ここで、L は f のキーワードであり、$f(L) = L'_i$ は f の値を意味する。

本論文で述べる実験では、以下の 4 種類の語彙関数を使用する：**Anti**、**Func**$_0$、**Oper**$_1$、および **Son**。以下に、それぞれの以上の語彙関数に対する意味と例を挙げる。

- **Anti** は反意語を表す。例：大きい——小さい、多い——少ない、泣く——笑う。
- **Func**$_0$ は軽動詞で「発生する」という意味を持

ち、文法的には主語のキーワードの述語として機能する。[4] 例：雨——降る、可能性——存在する、時間——飛び去る、疑問——生じる。

- **Oper**$_1$ は軽動詞（補助動詞）で、最初の意味的行為者（主語の主体役割）と状況の名称 (直接目的語) を関連付ける。例：風邪——ひく 、影響——与える 、結論——出す・導く・下す 。
- **Son** は、ある生物、物、出来事に対して、その特徴的若しくは表現的な音を示す。例：オオカミ——遠吠え、鳥——さえずり、機械——轟音。

2.2 実験データ

本研究で使用したデータセットには、中国語データセット 4 つと日本語データセット 4 つが含まれており、それぞれが前述の語彙関数のいずれかに対応している。各データセットは、複数の単語と、それに対応する上記の語彙関数の値を含む問題と解答集である。

中国語データセットは、現代中国語コーパスから抽出されており、日常生活、ニュース報道、文学作品など、さまざまな分野にわたる文を含んでいる。選定されたコーパスは、データの多様性と代表性を確保するためのものである。また、データセットの校正作業は『現代漢語詞典』（第 7 版）[5] と『漢語大詞典』[6] を基に行った。

日本語データセットは、日本国立国語研究所のコーパスから抽出され、口語会話、書面テキスト、ニュース報道など、多様な言語使用場面が含まれている。データセットの校正作業は『大辞林』[7] と『日本国語大辞典』[8] を基に行った。

これらの辞典に基づく校正によって、データセットの正確性と一貫性が保証され、各データは人工的に監査され、曖昧さによる精度の誤判定が防がれている。これにより、実験結果への影響が軽減された。

2.3 実験で使用した大規模言語モデル

複数の大規模言語モデルに語彙関数に関する類推関係問題を解答させた上、その回答を比較するため、実験では LangChain[1] を中心に使用し、統一インターフェースを設計した。LangChain を活用することで、同一の実験環境内で複数の大規模言語モデルを呼び出し、効率的にテストを管理、実行、記録することが可能となる。

1） https://www.langchain.com/

2.4 実験方法

本実験では、7 つの大規模言語モデルと 2 か国語における 4 つの語彙関数に関する問題セットを組み合わせ、最終的に 56 の独立したタスクを生成した。これらのタスクは実験内で「タスク」と呼び、各タスクは特定の大規模言語モデルに対して特定の語彙関数データセットを適用してテストを行う。タスク間の一貫性を確保するため、異なる語彙関数と言語に対して、あらかじめタスクのプロンプトテンプレートを設定した。これらのプロンプトは、大規模言語モデルに対して語彙関数の意味的な内包を説明するものである。また、プロンプトエンジニアリング [9] の few-shot prompting 手法を使用し、モデルの類推タスクにおける正確性向上を試みた。

語彙関数 **Anti** の日本語のタスクにおいて、大規模言語モデルに送信するプロンプトは 表 4 に示す。示されているプロンプトは、LangChain のテンプレートクラスによってテキストテンプレートがレンダリングされた結果であり、プロンプト内の例や質問はすべてプログラムがデータセットから自動的に取得し、追加したものである[2]。

表 1 プロンプトの例

構成要素	内容
身分提示語	あなたは類推推論タスクを得意とし
形式制限語	常に 1 つの単語だけを回答します。もし複数の答えがある場合は、最も一般的な答 えのみを回答します
例 1	例えば、“大” という質問に対する答えは次のようになります:“小”
例 2	例えば、“前” という質問に対する答えは次のようになります:“後”,“後ろ”
例 3	例えば、“高い” という質問に対する答えは次のようになります:“低い”
例 4	例えば、“明るい” という質問に対する答えは次のようになります:“暗い”
例 5	例えば、“広い” という質問に対する答えは次のようになります:“狭い”
質問	問題 “拡大” の答えは何ですか？

データセット内のパラメータとしての単語列を大

2） 実験プログラムでは、このプロンプトは実際には LangChain フレームワークのプロンプトクラスのインスタンスであり、リストに示された純粋なテキスト形式ではない。

規模言語モデルへの入力として使用し、プロンプトテンプレートを活用して完全なプロンプトを構築する。その上で、大規模言語モデルの出力形式を定義し、その出力をデータセット内の関数に対応する出力単語列と比較する。回答に正しい語彙関数の出力が含まれている場合、そのタスクに対する大規模言語モデルの回答を「正」と見なす。この方法を用いてすべてのタスクを個別に統計し、最終的にこれらのタスクの結果を集計して、さまざまな観点から分析を行う。また、実験結果は、モデルのサイズと語彙関数タスクのパフォーマンス間の相関関係を分析するために使用する。

3　実験結果と考察

本実験では、以下のモデルを使用して上記記述した実験を実施した：llama3.2:latest、llama3.2:1b、llama3.1:latest、phi3:latest、phi3:medium、gemma2:2b、gemma2:latest。語彙関数の出力は単語の集合であり、類推課題ではモデルが特定の意味関係を持つ単語や文を理解し生成する必要がある。さらに、多様な回答を得るために、過度に確信を持った出力を避ける目的で、各モデルの温度パラメータを 1 に設定した。

類推課題においてはモデルが特定の意味関係を理解し、それに基づいて適切な語彙や文を生成することが求められると考えている。温度を 1 に設定することで、モデルに一定の探索空間を与え、出力選択時の柔軟性を高め、複雑なタスクにおいてより適切な答えを見つけるのを助けることができると考えた。

3.1　タスクの正答率

表 2　実験の精度結果 (%)

	Anti		**Func$_0$**		**Oper$_1$**		**Son**	
モデル	ja	zh	ja	zh	ja	zh	ja	zh
llama3.2:latest	59	52	13	14	11	17	17	14
llama3.2:1b	27	46	8	11	6	7	1	12
llama3.1:latest	66	55	21	37	19	17	26	15
phi3:latest	41	37	14	15	9	15	5	8
phi3:medium	65	45	23	30	24	19	24	13
gemma2:2b	53	60	24	20	14	15	20	15
gemma2:latest	80	69	41	37	19	26	40	17

実験の精度結果 は表 2 に示す。直感的に理解しやすくするため、グループ別の棒グラフを用いて可視化した（図 1 と図 2）。実験結果を分析したところ、全体的なパフォーマンスの傾向として、用いたすべての大規模言語モデルが語彙関数 **Anti** のタスクで高い正答率を示し、その正答率は通常 50%から 70%の範囲に収まっていた。このタスクは、比較的簡単な反義語関係に基づくものであり、モデルが適切に処理しやすいことが分かる。

一方、語彙関数 **Func$_0$** や **Son** のタスクでは、正答率は中程度で、通常 20%から 40%の範囲に収まった。これらのタスクは、語彙関数の意味関係を正確に理解し生成する必要があるため、モデルにとってやや難易度が高く、正答率が低くなった。また、その他の語彙関数のタスクでは、さらに低い正答率が観察される傾向があった。

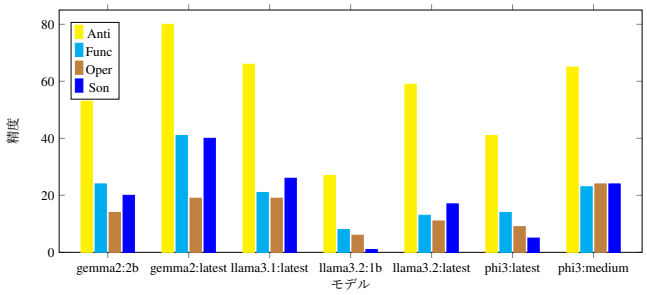

図 1　タスク精度の統計グラフ（日本語）

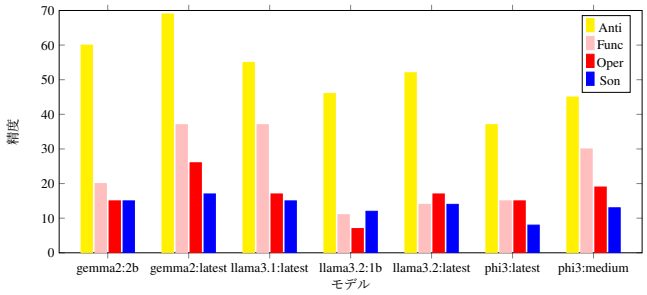

図 2　タスク精度の統計グラフ（中国語）

また、モデル別のタスク結果を観察すると、異なる語彙関数間での正答率の差はほとんど均等であり、すべてのモデルにおいて、反義語の理解と出力に関するタスクを他の語彙関数のタスクよりも容易に処理していることが示されている。このことから、反義語関係を処理するタスクは、他の語彙関数のタスクと比較して難易度が低いと考えられる。

モデル別に見ると、gemma2 はほとんどすべてのタスクで高い正答率を示し、平均で約 40%の正答率を記録した。一方、llama3.2:1b はほぼすべてのタスクで他のモデルよりも低い正答率を示し、その平均正答率は約 10%にとどまった。この結果は、モデルの規模や訓練データの質、または特定のタスクへの適応能力の違いを反映している可能性がある。

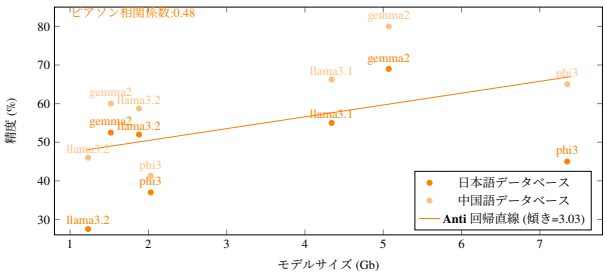

図 3　語彙関数 **Anti** の相関散布図

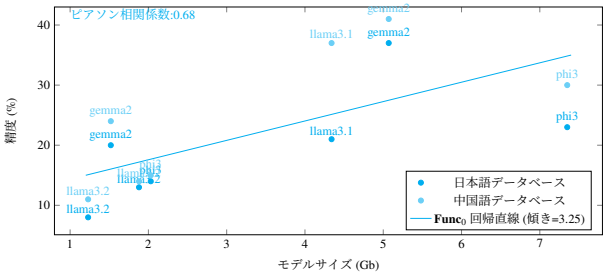

図 4　語彙関数 $\mathbf{Func}_0$ の相関散布図

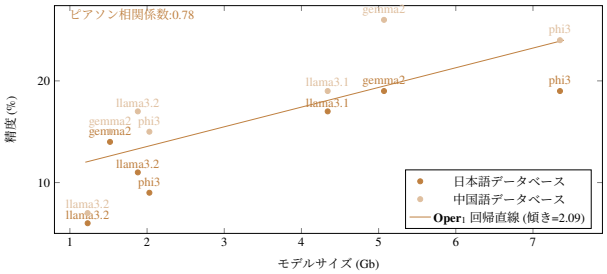

図 5　語彙関数 $\mathbf{Oper}_1$ の相関散布図

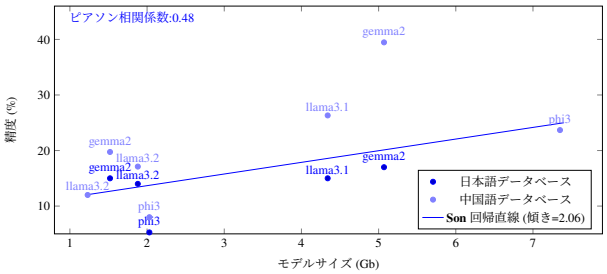

図 6　語彙関数 **Son** の相関散布図

3.2　モデルのサイズとタスク精度の相関分析

以上のモデルごとの正答率の差異を明らかにするため、正答率とモデルサイズの相関関係を分析した。その結果を示す散布図は、図 3、図 6、図 4、および図 5 に示す。相関性分析の結果、すべてのタスクにおける総合的な傾向として、$\mathbf{Oper}_1$ の相関係数が最も高く、Pearson の相関係数は 0.78 であることが分かった。しかし、その精度は 4 つの語彙関数の中で最も低い結果となった。一方、**Anti** および **Son** の相関係数は最も低かったが、**Anti** の精度は最も高い結果を示した。

これらの難易度の考察に基づいて解釈すると、難易度が高いタスクでは相関性が高く、難易度が低い **Anti** のタスクでは相関性が低い傾向にあることがわかった。これは、**Anti** のタスクの難易度が特定の分岐点の下にあり、他のタスクはその分岐点を超えていることを示唆している可能性がある。分岐点を超えると、タスクの難易度、相関性が高くなる一方、分岐点以下のタスクでは相関性が低くなる傾向が見られた。

4　実験の限界と今後の改善点

上述の実験からいくつかの明確な結論が得られたが、同時にいくつかの課題も浮き彫りになった。

まず、実験に使用したモデルの数が限られている点である。本研究では 7 つの大規模言語モデルを用いて実験を行ったが、この数では相関性の計算に十分でない可能性がある。モデルの種類を増やし、タスクに参加するモデルの数を増やすことで、より安定した相関係数の結果を得ることが期待できる。現在の結果では、同じファミリーに属するモデル間で精度がモデルサイズと相関している傾向が顕著である。このため、モデル数を増やす際には、同一ファミリーに属するモデルの数を増やし、ファミリー内での追加的な分析を行うことが有益である。

次に、実験で使用したデータセットにも改善の余地がある。本実験で用いたデータセットは様々な資料から手動で抽出したものであり、一部の語彙関数は複雑な意味を持つため、データセット内に不正確な用例が含まれ、回答に漏れが生じることがある。このため、誤った判断が下され、実際の正答率よりも低い精度が得られることがあると思われる。データセットの改良を進め、より精緻な注釈と完全な回答を提供することで、実験結果の信頼性を向上させることが可能となる。

5　まとめ

本論文では、語彙関数データセットの概念に基づいた類推関係問題タスクを用い、複数の大規模言語モデルにおいて、日本語と中国語で単語間関係の知識程度を評価した。比較的簡単な反義語関係の語彙関数に対して、難易度の高い語彙関数を分かった。また、大規模言語モデルのサイズと正確率の相関性と語彙関数の難易度の関係を解析してみた。

参考文献

[1] Sébastien Marengo (sous la direction de). **La théorie sens-texte**. L'Harmattan, Paris, 2021.

[2] Alain Polguère Igor Mel'čuk. **Les fonctions lexicales dernier cri**. L'Harmattan, 2021.

[3] Thierry Fontenelle. Using a bilingual dictionary to create semantic networks. **Practical Lexicography: A reader**, pp. 175–185, 2008.

[4] Igor A. Mel'cuk. Lexical functions in lexicographic description. **Proceedings of the Eighth Annual Meeting of the Berkeley Linguistics Society**, pp. 427–444, 1982.

[5] Dictionary Editorial Office of the Institute of Linguistics. **A Modern Chinese Dictionary**. The Commercial Press, Chinese Academy of Social Sciences, 7th edition, 2016.

[6] Luo Zhufeng. **The Great Chinese Dictionary**. The Great Chinese Dictionary Press, 1986.

[7] 松村明. 大辞林第. 三省堂, 三版, 2006.

[8] 日本大辞典刊行会. 日本国語大辞典. 小学館, 第二版, 2002.

[9] DAIR.AI. Prompt engineering guide. https://www.promptingguide.ai/, 2024.

A 付録

A.1 プロンプト

SystemMessage	あなたは類推推論タスクを得意とし、常に 1 つの単語だけを回答します。もし複数の答えがある場合は、最も一般的な答えのみを回答します
HumanMessage	例えば、”狼”という質問に対する答えは次のようになります:
AIMessage	うなる, 吠える, 吼く
HumanMessage	例えば、”オオカミ”という質問に対する答えは次のようになります:
AIMessage	うなる, 吠える, 吼く
HumanMessage	例えば、”狐”という質問に対する答えは次のようになります:
AIMessage	なく
HumanMessage	問題”トラ”の答えは何ですか？
AIMessage	唸る

A.2 データベースの例

表 3 日本語での **Son** の例

入力	出力
狼	うなる、吠える、吼く
オオカミ	うなる、吠える、吼く
狐	なく
トラ	うなる、吠える、吼く
ライオン	うなる、吠える、吼く
鳥	鳴く、さえずる
小鳥	鳴く、さえずる
猫	鳴く
カエル	鳴く
コオロギ	鳴く、さえずる
セミ	鳴く、さえずる
牛	鳴く
羊	鳴く
馬	鳴く
鹿	鳴く
サル	鳴く、叫ぶ、わめく
チーター	鳴く
豚	鳴く
鐘	鳴る、なる
チャイム	鳴る
ベル	鳴る、響く
太鼓	鳴る
笛	鳴る

表 4 中国語での **Anti** の例

入力	出力
大	小
前	后
高	低
明亮	昏暗
新	旧
多	少
生	死
笑	哭
正面	反面
上	下
男	女
天	地
前面	后面
死	活
希望	失望