指標比喩コーパスにおける印象評定情報を用いた LLM の指標比喩の理解能力評価

加藤真輝 [1]　古宮嘉那子 [1]　加藤祥 [3]　宇野良子 [1]　浅原正幸 [2]
[1] 東京農工大学 [2] 国立国語研究所 [3] 北海道大学
{s243217z@st.go, kkomiya@go, ryokouno@cc}.tuat.ac.jp,
katosachi@let.hokudai.ac.jp, masayu-a@ninjal.ac.jp

概要

比喩表現は円滑なコミュニケーションや言語表現の拡張など，さまざまな場面で用いられる．一方で，一般話者の比喩に対する感覚を LLM がどの程度捉えられるかを定量的に分析した研究は限られている．そこで本研究では，指標比喩の用例に一般話者の「比喩性」「わかりやすさ」の印象評定が付与されたデータを用い，評定値を自動予測するモデルを構築した．実験の結果，印象評定との相関係数は 0.34–0.75 であった．また，被喩辞・喩辞の情報を明示的に入力に加えることで，比喩性の予測性能が向上していることを確認した．

1　はじめに

比喩表現とは，ある領域の概念構造を別の領域の理解に利用する表現である．例えば「雪のような肌」という用例では「雪」が喚起する白さや綺麗さといったイメージを「肌」に適用することで，肌の状態を具体的に描写できる．このような比喩表現は，円滑なコミュニケーションや言語表現の拡張など，さまざまな場面で用いられる．

多様な自然言語処理タスクにおいて高い性能を示す大規模言語モデル（LLM）は，このような比喩を含む用例も大量に学習しているため，比喩を一定程度理解していると考えられている．一方で，一般話者がもつ比喩に対する直感的な感覚を，LLM がどの程度捉えられるのかを分析した研究は限られている．

本研究では，加藤らによって作成された指標比喩の用例が収集されたコーパスである BCCWJ-Simile[1, 2] と，中村による『比喩表現の理論と分類』[3] の，加藤らにより電子化されたデータ [4] の一部を用いて，LLM が一般話者の比喩理解をどの程度持っているか評価した．これらのコーパスには一般人が持つ一般話者が抱く「比喩性」に対する印象評定がクラウドソーシングにより収集されている．そこで本研究では，この印象評定値をどれだけ正確に予測できるかを指標として，LLM の比喩理解能力を測定した．さらに，同じデータには「わかりやすさ」に関する印象評定も含まれているため，比喩の理解難易度をモデルが把握できるかについても併せて評価した．予測モデルには，日本語 ModernBERT[1)] および OpenAI 社が提供する文章埋め込みモデル text-embedding-3-large[2)] を用いた．

2　関連研究

直喩（指標比喩とほぼ同義）に関する自然言語処理研究として，鈴木ら [5] は「S のような T」という形式の直喩が，字義的に S と T を同一視する解釈と，比喩として理解した上での解釈とで異なる点に着目している．この問題意識に基づき，S と T が与えられたときに，その直喩から導かれる解釈表現の集合を生成するタスクを提案・検討した．例えば「ひまわりのような笑顔」に対しては，「明るい」「暖かい」といった形容詞が生成されることが期待され，直喩の解釈生成を扱う枠組みを提示している．

また，蜜石ら [6] は，認知心理学における比喩の処理過程を工学的手法に取り入れ，比喩の適切さを計算する手法を提案した．さらに，直喩生成のためのデータセット構築と，適切な直喩表現を自動選択するモデルの学習を行い，比喩表現の適切さ評価・生成を実証的に扱っている．

一方で，我々の知る限り，クラウドソーシング等により大規模に収集された一般話者による印象評定を直接の教師信号として，言語モデルの学習や性能

1） https://huggingface.co/sbintuitions/modernbert-ja-310m
2） https://platform.openai.com/docs/models/text-embedding-3-large

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

分析に取り組んだ研究は存在しない．

3 データ

本研究では比喩のうち指標比喩を対象として分析を行った．

3.1 指標比喩

中村 [3] は，比喩性の表現あるいは把握における何らかの言語形式である「比喩指標」を含む比喩表現を「指標比喩」と分類している（直喩ともいう）．例えば「雪のような肌」は「ような」という比喩指標を含んでおり，「肌」を「雪」で喩えている．また，比喩においては一般に例えられる語を被喩辞，例える語を喩辞と呼ぶ．「雪のような肌」の例では被喩辞は肌であり，喩辞は雪である．

3.2 コーパス

本研究では，指標比喩用例が収集されたコーパスである BCCWJ-Simile[1] と，『比喩表現の理論と分類』[3] の電子化データ [4] の一部を用いて実験を行った．

3.2.1 BCCWJ-Simile

BCCWJ-Simile は，現代日本語書き言葉均衡コーパス（BCCWJ）[3)] のコアデータから指標比喩の用例を収集したコーパスである [1, 2]．本コーパスは以下手順により作成されている．

1. 手掛かり句の認定と検索
   「まるで」「のような」などの指標比喩であることを示す手がかり句を用い，BCCWJ のコアデータから検索を行い文意を確認する．
2. MIP に基づく比喩認定
   MIP[7] は，対象語の文脈を踏まえた語義が，対象語の基本義・代表義と逸脱している場合に比喩と認定する手法であり，本手法を用いて検索を行った用例に指標比喩が含まれているか確認する．指標比喩と認定された場合，喩辞と被喩辞の同定も行う．
3. 印象評定の収集
   3.3 で述べる印象評定の収集も行われた．

本研究では，コーパス内の各用例と対応する被喩辞，喩辞，印象評定値を利用した．

3） https://clrd.ninjal.ac.jp/bccwj

3.2.2 『比喩表現の理論と分類』データ

中村 [3] は，明治〜昭和期の文学作品 50 種を対象に比喩表現を収集し，比喩に関する情報を抽出した．本研究では，このうち指標比喩に該当するデータを扱う．指標比喩については，加藤ら [4] により，BCCWJ-Simile と対照可能な形式で電子化・整備されたデータが提供されているため，これを利用した．具体的には，BCCWJ-Simile と同様に，コーパス内の各用例と対応する被喩辞，喩辞，印象評定値を利用した．

3.3 印象評定

加藤ら [4] では指標比喩の用例に対し，Yahoo!クラウドソーシングを用いて印象評定値が収集されており，本研究ではこれを一般話者における評定として扱った．収集項目は「比喩性」「新奇性」「わかりやすさ」「擬人化」「具体化」であり，「0: 全く違う」〜「5: そう思う」の 6 種類の評定値から 1 つ選択される．本研究では「比喩性」と「わかりやすさ」の評定値を予測対象として利用した．比喩性の収集では「何かを他の物事でたとえる（比喩）表現を含むか」という質問を行い，わかりやすさの収集では「わかりやすく表現されていると思いますか」という質問を行った上で用例が提示された．用例は，比喩指標のマークアップ等の操作は行わず提示された．

4 手法

4.1 ModernBERT を用いた印象評定予測モデル（MB）

本研究では，SB Intuitions 株式会社が開発した日本語 ModernBERT を用いる．ModernBERT[8] は，BERT[9] の設計を近代化したエンコーダ型 Transformer であり，RoPE による長文対応（最大 8192 トークン）と，Unpadding や FlashAttention 等による計算効率化を特徴とする．本モデルは日本語を含むデータで事前学習されている．

印象評定値の予測では，入力特徴量として用例・喩辞・被喩辞を組み合わせる．各特徴量を独立に日本語 ModernBERT へ入力し，最終層における先頭トークン（[CLS] 相当）のベクトル表現を抽出する．抽出したベクトルは連結（concatenation）し，回帰予測用の多層パーセプトロン（MLP）に入力することで，出力として印象評定値を得る．

ModernBERT のパラメータは学習可能である一

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 1　「比喩性」予測モデル構築の設定（MB: 4.1 で定義したモデル，4.2 で定義したモデル）

モデル	コーパス	隠れ層の次元	隠れ層の数	学習率	エポック数
MB	BCCWJ-Simile	1024	1	0.0005	15
MB	比喩表現の理論と分類	1024	1	0.001	5
OA	BCCWJ-Simile	512	2	0.005	8
OA	比喩表現の理論と分類	1024	3	0.005	10

表 2　「わかりやすさ」予測モデル構築の設定（MB: 4.1 で定義したモデル，4.2 で定義したモデル）

モデル	コーパス	隠れ層の次元	隠れ層の数	学習率	エポック数
MB	BCCWJ-Simile	1024	1	0.0005	10
MB	比喩表現の理論と分類	1024	2	0.0005	10
OA	BCCWJ-Simile	512	2	0.005	5
OA	比喩表現の理論と分類	1024	3	0.005	10

方，本研究では事前学習済み表現の有効性を評価するため，ModernBERT 本体のパラメータ更新は行わず固定し，MLP のみを学習した．

また，入力特徴量の寄与を検証するため，用例・喩辞・被喩辞を入力とした場合と用例のみを入力とした場合のモデルを学習した．それぞれ入力の次元が異なるモデルとなるが，学習のハイパーパラメータは同一とした．以後，本節で定義した ModernBERT を利用したモデルを MB と呼ぶ．

4.2　text-embedding-3-large を用いた印象評定予測モデル（OA）

text-embedding-3-large は，OpenAI 社が提供する文章埋め込みモデルであり，API を通じて利用可能である．まず，用例・喩辞・被喩辞それぞれについて埋め込みベクトルを取得する．次に各ベクトルを連結し，MLP へ入力して印象評定値を予測する．

学習手順および評価手順は ModernBERT の場合と同一とし，損失関数に MSE を用いる．また，同様に用例のみを入力としたモデルも学習した．なお，text-embedding-3-large は API から得られる固定表現のみ利用可能であるため，本モデルの学習対象は MLP 層のみである．以後，本節で定義したモデルを OA と呼ぶ．

5　実験

ハイパーパラメータは，入力特徴量を「用例+喩辞+被喩辞」とした状態でグリッドサーチにより選定した．比喩性予測モデルで採用した値を表 1 に，わかりやすさ予測モデルで採用した値を表 2 に示す．学習率は，選定した初期値を基準としてコサインスケジューリングにより段階的に減衰させた．

損失関数には平均二乗誤差（MSE）を用い，最適化手法には Adam を採用した．過学習を抑制するため，早期終了（early stopping）を導入した．具体的には，検証損失の最良値更新が 0.001 未満の場合を改善なしとみなし，その状態が 3 エポック連続した時点で学習を終了した．加えて，ドロップアウト率は 0.1 とした．

実験は 5 分割交差検証により実施し，各バッチ内で z 標準化を行って学習した．使用データ数は，BCCWJ-Simile が合計 827 件，『比喩表現の理論と分類』データが合計 899 件であった．バッチサイズは 16 とした．

6　評価と考察

評価指標として，テストセットに対する平均絶対誤差（MAE），二乗平均平方根誤差（RMSE），Pearson 相関係数 r を算出した．MAE は，予測値と真値の差の絶対値をとり，その平均として求める．RMSE は，予測値と真値の差を二乗し，その平均をとったうえで平方根を取って求められる．各指標は元データの尺度（0–5）で計算した．

5 分割交差検証を行い，各フォールドで算出した評価指標の平均を表 3, 4, 5, 6 に示す．これらの表は，BCCWJ-Simile または『比喩表現の理論と分類』データを用いた場合の，「比喩性」または「わかりやすさ」予測結果に対応する．なお，表 3, 4, 5, 6 に示すすべての相関係数 r について，各フォールドで得られた相関の有意性を検定し，フォールド間の p 値を Fisher の結合確率検定で統合した結果，すべて $p < 0.001$ であった．

BCCWJ-Simile の比喩性予測では相関係数 r が最大 0.750，わかりやすさ予測では最大 0.597 であった．このことから，LLM を用いて一般話者の印象評

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 3 BCCWJ-Simile, 比喩性予測の評価

モデル	入力特徴量	MAE	RMSE	Pearson r
MB	用例	0.718	0.886	0.538
MB	用例+被喩辞+喩辞	0.621	0.775	0.698
OA	用例	0.635	0.802	0.638
OA	用例+被喩辞+喩辞	0.541	0.679	0.750

表 4 BCCWJ-Simile, わかりやすさ予測の評価

モデル	入力特徴量	MAE	RMSE	Pearson r
MB	用例	0.367	0.463	0.597
MB	用例+被喩辞+喩辞	0.383	0.477	0.549
OA	用例	0.387	0.501	0.486
OA	用例+被喩辞+喩辞	0.392	0.507	0.463

表 5 比喩表現の理論と分類, 比喩性予測の評価

モデル	入力特徴量	MAE	RMSE	Pearson r
MB	用例	0.236	0.306	0.18
MB	用例+被喩辞+喩辞	0.236	0.293	0.344
OA	用例	0.252	0.315	0.25
OA	用例+被喩辞+喩辞	0.241	0.306	0.295

表 6 比喩表現の理論と分類, わかりやすさ予測の評価

モデル	入力特徴量	MAE	RMSE	Pearson r
MB	用例	0.239	0.293	0.436
MB	用例+被喩辞+喩辞	0.244	0.301	0.376
OA	用例	0.243	0.300	0.358
OA	用例+被喩辞+喩辞	0.238	0.298	0.341

定を一定の相関で再現できることが確認された．一方,『比喩表現の理論と分類』データでは，比喩性予測の相関係数が最大 0.344，わかりやすさ予測では最大 0.436 となり，BCCWJ-Simile より相関が低かった．この差は，BCCWJ-Simile が一般的な書き言葉を広範囲に収集したデータであるのに対し,『比喩表現の理論と分類』データは文学作品の用例を中心とするため，解釈に必要な背景知識が多いことや文脈理解が複雑であることに起因する可能性がある．

モデル間の違いについては，BCCWJ-Simile の比喩性予測では OA を用いたモデルが最も高い相関を示した一方,『比喩表現の理論と分類』データの比喩性予測では MB モデルが最良であるなど，一貫した優位性は確認できなかった．この結果は，学習モデルに改善の余地があることに加え，用例が属するドメインや予測対象（比喩性／わかりやすさ）によって適切なモデルが変化しうることを示している．

入力特徴量の違いについて，比喩性の予測では用例+被喩辞+喩辞を入力とした場合の方が用例のみを入力とした場合より相関係数が一貫して高かった．これは，比喩性の印象予測が「何を何に見立てるか」という対応づけ情報の明示に強く関係する可能性を示す．一方，わかりやすさの予測では，両者で差がないもしくは性能が低下する傾向が見られた．このことから，一般話者の直感的なわかりやすさの予測においては，被喩辞と喩辞が与える比喩の対応関係情報が必ずしも有効ではなく，場合によっては冗長となる可能性がある．

7 課題

本研究の課題として，まず『比喩表現の理論と分類』データにおける予測性能が相対的に低い点が挙げられる．この点は，データ量の拡充やモデルの変更等により性能向上を図れる可能性がある．

また，本研究で用いたコーパスは，指標比喩用例に対して喩辞と被喩辞が手動アノテーションされたものである．しかし，喩辞・被喩辞のアノテーションが付与されない状況では，比喩性予測で最良の性能を示した用例+被喩辞+喩辞を入力とするモデルを適用できない．この点は，喩辞・被喩辞の自動推定モデルの学習や，喩辞・被喩辞に依存しないモデルアーキテクチャの構築により解決が期待される．

8 おわりに

本研究では，LLM を利用して指標比喩用例に関する比喩性とわかりやすさを一定の性能で予測できることを示し，LLM が一定の比喩理解能力を有することを示した．今後の研究の方向として，本研究では指標比喩の用例を対象としたが，比喩表現には隠喩や文脈比喩などさまざまな種類があり，他の比喩表現に対しても一般話者の印象評定を予測していくことが考えられる．また，今回は「比喩性」および「わかりやすさ」に着目したが，他の評価項目に拡張することも可能である．例えば浅原ら [10] は新たに「文学性」の印象評定を収集しており，それに着目することで，LLM の比喩理解能力に関する分析をさらに進められる可能性がある．

また，蜜石ら [6] は，認知心理学における比喩の処理過程を取り入れ，比喩の適切さを計算する手法を提案している．本研究においても認知心理学の知見を工学的手法に取り入れることで，LLM の比喩理解能力向上への寄与と，人間の比喩理解研究への貢献を目指すことができると考える．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は JSPS 科研費 JP22K12145, JP25K00459, JP25K15839，及び国立国語研究所共同研究プロジェクト「アノテーションデータを用いた実証的計算心理言語学」の助成を受けたものです.

参考文献

[1] 加藤祥, 菊地礼, 浅原正幸. 『現代日本語書き言葉均衡コーパス』に基づく指標比喩データベース. 自然言語処理, Vol. 27, No. 4, pp. 853–887, 2020.

[2] 菊地礼. 日本語指標比喩のデータセットの紹介とその活用. （JLR2024 発表資料）, March 2024.

[3] 中村明. 比喩表現の理論と分類. 秀英出版, 1977.

[4] 加藤祥, 浅原正幸. 『比喩表現の理論と分類』データの電子化および情報付与. 国立国語研究所論集, Vol. 25, No. 4, pp. 1–19, 2023.

[5] 鈴木颯仁, 山田寛章, 徳永健伸. 事前学習済みモデルを用いた日本語直喩表現の解釈. 言語処理学会 第 30 回年次大会 発表論文集, 2024.

[6] 蜜石湧斗, 嶋田和孝. 直喩生成データセットを用いた比喩表現の自動選択. IEICE Technical Report NLC2018-42, The Institute of Electronics, Information and Communication Engineers (IEICE), 2019.

[7] Pragglejaz Group. Mip: A method for identifying metaphorically used words in discourse. **Metaphor and Symbol**, Vol. 22, No. 1, pp. 1–39, 2007.

[8] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics.

[9] 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] 浅原正幸, 加藤真輝, 古宮嘉那子, 加藤祥. テキストの文学性の評価について. 言語処理学会 第 32 回年次大会 発表論文集, 2026.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).