大規模言語モデルで生成された文学テキストの検出と有効な識別的特徴の探求

岩本海風　宮本友樹　内海彰
電気通信大学
i2430010@edu.cc.uec.ac.jp　{miyamoto,utsumi}@uec.ac.jp

概要

本研究は，人間が作成した文学テキストと LLM が生成した文学テキストの自動分類がどの程度の精度で可能かを BERT モデルを用いて検証した．その結果，高精度で可能であることが示された．また，人間と LLM の生成テキストの間に存在する識別可能な特徴を探求するため，Random Forest で 18 種類の特徴量を用いて LLM 生成テキストの検出を行った．その結果，詩においては，MATTR，Lexical Density，文長のばらつき，感情の揺らぎ，歌詞においては，Maas Index，Hapax Legomenon Ratio，文長のばらつき，感情の揺らぎ，短編小説では，HD-D，隣接文類似度，文長のばらつき，修飾関係の意味的距離が分類に寄与していることが示された．

1　はじめに

近年，大規模言語モデル（LLM）は目覚ましい進歩を遂げており，ニュース記事や学術論文といった一般的なテキストに加え，詩や小説のような文学テキストも生成可能になっている．文学テキストは，創造性や独自の文体といった非定型的な要素を含む点で一般的なテキストと異なる特徴を持つ．しかし，LLM が生成した文学テキストと人間が書いた文学テキストの区別は困難になっている．Porter らの研究 [1] では，読者による LLM 生成詩の特定で正答率が 46.6 %となり，LLM 生成詩を実際の人間が書いた詩よりも人間が書いたものと判断する傾向が強いことが示された．人手による LLM の生成した文章の識別は困難であるが，この困難さはコンピュータを用いた場合にも当てはまるとは限らない．本研究では，人間が作成した文学テキストと LLM が生成した文学テキストの自動分類がどの程度の精度で可能かを検証する．特に，高い検出性能が示された場合には，人間と LLM が生成した文学テキストの間に存在する識別可能な特徴を解明する．LLM が生成した文学テキストを検出できれば，教育面において，学生が自分の力で文章を作成することを促し，思考力や創造性を守ることに繋がる．

2　関連研究

LLM が生成した文章の検出に関する先行研究 [2, 3, 4] は少なくない．これらの研究では，LLM が生成した一般的なテキストの自動検出は高性能であることが示されているが，LLM が生成した文学テキストの検出においては物語文の検出が高性能であることしか示されていない．また，Hamat らの研究 [5] では古典詩を対象に LLM が生成した文章の検出をしているが，LLM が古語的表現に十分に対応できていない可能性があるため，この結果は現代詩には直接適用できない可能性がある．さらに，LLM 生成詩が単純なプロンプトで作成されており，人間作成詩との間で文字数や内容に乖離が生じている可能性がある．本研究は，先行研究の課題を解決するため，LLM が生成した文章の検出対象として現代の文学テキスト，形式としては詩，歌詞，俳句，短編小説に焦点を当てる．また，データセットの作成において，人間が生成した文学テキストと LLM が生成した文学テキストの文字数や内容に差異が生じないようにプロンプトを作成する．

3　データセットの作成

分類器を構築するため，人間または LLM が生成した文学テキストのデータセットを作成する．

人間が作成した文章　詩と俳句は，インターネット上にある『ポエム投稿サイト｜詩人たちの小部屋』，『日本俳句研究会／ IT 俳句会』からそれぞれ収集する．歌詞は株式会社シンクパワーが提供する「歌詞コンテンツデータ集」，短編小説は青空文庫で公開されているものから収集する．収集する詩，歌

表 1 LLM で生成するときに使用するプロンプト

種類	プロンプト
詩 1	あなたは詩人です．"title"をテーマにして，○○文字以上〜○○文字以内で詩を生成してください．
詩 2	あなたは詩人です．"word1","word2","word3"の単語を使って，○○文字以上〜○○文字以内で詩を生成してください．
歌詞 1	あなたは作詞家です．"title"をテーマにして，○○文字以上〜○○文字以内で歌詞を生成してください．
歌詞 2	あなたは作詞家です．"word1","word2","word3"の単語を使って，○○文字以上〜○○文字以内で歌詞を生成してください．
俳句	あなたは俳人です．"kigo"の単語を使って，俳句を 1 つだけ生成してください．
小説	あなたは作家です．次の要約をもとに，○○文字以上〜○○文字以内で短編小説を生成してください．"[要約文]"

詞，俳句，短編小説は，それぞれ 1600 件である．

LLM が生成した文章 Gemini2.5, GPT-4o, Claude-3，Deepseek-V3 のそれぞれの API を利用し，詩，歌詞，俳句，短編小説を生成する．各文学テキストの形式ごとに，4 種類の LLM のそれぞれで 400 件，合計で 1600 件の文学テキストを生成する．

LLM による生成では，人間が作成した文学テキストと同じ長さや内容になるように，表 1 に示すプロンプトを与える．文字数については，人間が作成したテキストの文字数を 10 の位で切り捨て，そこから 100 文字増やした範囲で生成するようにプロンプトで指示する．詩または歌詞の生成では 2 種類のプロンプトを扱い，各プロンプトごとに，4 種類の LLM のそれぞれで 200 件，合計 800 件を生成する．"title"には人間が作成した詩または歌詞の題名，"word1"，"word2"，"word3"には人間が作成した詩または歌詞における頻度上位 3 単語が入る．また，俳句の生成で使われる"kigo"には人間が作成した俳句で使われている季語が入る．そして，短編小説の生成では人間が作成した短編小説を 100〜200 文字以内になるように GPT-4o で要約した文章をプロンプトにおける [要約文] として扱う．

4 LLM 生成テキストの検出

4.1 分類器の構築

東北大学の日本語事前学習済みの BERT モデルをファインチューニングし，人間が作成した文学テキストと LLM が生成した文学テキストを識別する分類器を作る．エポック数は 3，学習率は 2e-5，バッチサイズは 16 で学習を行う．BERT モデルの一度に入力できる文章の長さの上限を超えた場合，文章を 512 トークンずつチャンクとして分割し，文章の繋がりの消失を防ぐために，128 トークンだけチャンク同士が重なるようにする．全てのチャンクの予測結果の平均を取り，文章全体の予測を決定する．

表 2 BERT の分類器による分類評価

テキストの種類	正解率	適合率	再現率	F1 スコア
詩	0.964	0.955	0.974	0.964
歌詞	0.977	0.966	0.988	0.977
俳句	0.856	0.808	0.936	0.867
短編小説	1.000	1.000	0.999	1.000

4.2 BERT による分類結果

文学テキスト形式別に学習およびテストを行い，5 分割交差検証を用いて算出した正解率，適合率，再現率，F1 スコアのマクロ平均を表 2 に示す．エッセイ，物語文，ニュース記事を LLM が生成した文章の検出対象としている He ら [4] の研究では，すべてのテキストの形式において，BERT 分類器の F1 スコアが 0.9 を超えている．表 2 によると，俳句以外の文学テキストでも正解率と F1 スコアが 0.9 を超えている．したがって，LLM が生成した一般的なテキストの検出と同様に，文学テキストの検出も高性能で可能であることがわかる．俳句が他の文学テキストの形式に比べて検出精度が劣るのは，俳句の形式が五七五であり，字数が限られているからであると考えられる．

5 LLM 生成テキストの検出に有効な特徴

4.2 節で，BERT をファインチューニングした分類器を利用することで，LLM が生成した文学テキストの検出が高性能で可能であることが示された．しかし，人間と LLM が生成した文学テキストを区別するような具体的な特徴は明らかになっていない．そこで，具体的な特徴量を使って分類を行い，どのような特徴量が LLM の生成する文学テキストの検出に寄与しているかを調べる．

5.1 語彙に関する特徴量

Hamat ら [5] は，Maas Index，MTLD，MATTR，HD-D，Hapax Legomenon Ratio，Lexical Density の 6 つの語彙指標（詳細は付録 A を参照）を用いて，LLM が

表 3　語彙指標を利用した分類評価

テキストの種類	正解率	適合率	再現率	F1 スコア
詩	0.780	0.780	0.780	0.780
歌詞	0.777	0.778	0.777	0.777
俳句	0.558	0.559	0.558	0.555
短編小説	0.901	0.901	0.901	0.901

生成した詩と人間が書いた詩の語彙的な豊かさを比較している．これらの指標は，テキストにおける語彙の多様性を捉える能力があると言われている [6]．Random Forest を用いて，人間が作成した文学テキストと LLM が生成した文学テキストを識別する分類器を作り分類を行う．アンサンブルを行う決定木の数を 100 に設定し学習を行う．

5.2　語彙指標を用いた分類結果

語彙指標を特徴量として分類を行い，5 分割交差検証で算出した正解率，適合率，再現率，F1 スコアのマクロ平均を表 3 に示す．表 2 で示された BERT による分類結果と比較すると，どの文学テキストも精度が劣っている．そのため，人間が作成する文学テキストと LLM が生成する文学テキストの間には，語彙以外の特徴もあると考えられる．

5.3　語彙以外の指標

今度は，以下に示す語彙以外の指標（詳細は付録 B を参照）も特徴量として加え，LLM が生成した文章の検出を行う．

文長のばらつき　文長の変化に関する指標である．1 つの文ごとに文字数を数え，分散を求める．

音素の美的スコア　音素の美的魅力に関する指標である．言葉の音構成に関する研究 [7] によると，一般的に滑らかで継続的な音は美しいとされ，突発的で荒い音は美しくないとされている．テキスト中の音素列を分析し，美的価値の高い音素と低い音素の出現割合に基づいて美的スコアを算出する．

修飾関係の意味的距離　分布意味論に基づいた指標である．文法的に強い関係（主述関係，名詞修飾，形容詞修飾，連用修飾）のあるペアについてコサイン類似度を求める．そして，1 からペアのコサイン類似度を引いた値の平均，分散，範囲（最大値と最小値の差）をそれぞれ求め，これら 3 つを特徴量とする．

具象性スコア　単語が具体的かどうかを評価する指標である．奈良先端科学技術大学院大学が公開している日本語抽象度辞書 [8] を用いる．辞書のスコアに基づき，辞書にヒットした単語のスコア合計を求める．そして，その合計に対して，ヒットした単語の総数で割った値が具象性スコアとなる．

表 4　全ての指標を利用した分類評価

テキストの種類	正解率	適合率	再現率	F1 スコア
詩	0.848	0.848	0.848	0.848
歌詞	0.875	0.876	0.875	0.875
短編小説	0.940	0.941	0.940	0.940

隣接文の類似度　文脈変化の多様性に関する指標である．Sentence Transformer[1] を用いて，それぞれの文をベクトル化し，隣接文同士のコサイン類似度を求める．そして，類似度の平均と分散と範囲を求め，これら 3 つを特徴量とする．

感情の揺らぎ　テキストにおける感情の変化に関する指標である．ポジティブ／ネガティブのスコアを出せるモデル[2] を用いて，それぞれの文ごとに感情スコアを計算する．そして，感情スコアの分散と範囲をそれぞれ求め，この 2 つを特徴量とする．

五感表現の割合　五感に関する表現が使われているかどうかを評価する指標である．日本語 WordNet を利用し，五感（視覚，聴覚，嗅覚，味覚，触覚）の各モダリティにおいて，知覚動作，感覚器，感覚特性に対応する最上位の単語を選定し，それらの関連語と合わせて五感語辞書を作成する．そして，五感表現の出現割合を計算する．

5.4　全ての指標を用いた分類結果

全ての指標を用いて分類し，5 分割交差検証で算出した正解率，適合率，再現率，F1 スコアのマクロ平均を表 4 に示す．なお，俳句については，文長や語数が制約されており，語彙以外の指標が適用できないため分類を行っていない．詩，歌詞，短編小説の 3 つの文学テキストの形式において，語彙指標のみを用いた場合よりも，語彙以外の指標も加えた場合の方が検出精度が高いことが示された．

また，分類における特徴量の重要度と，Cliff's Delta による効果量を表 5 に示す．なお，特徴量の重要度については，全データで学習した場合の重要度を示しており，重要度が 0.06 を超えている値を太字で示す．また，Cliff's Delta は，対象とする特徴量に対して，片方の群がもう片方の群よりも大きくなるデータペアの割合を示したものであり，正の値であれば人間の方が特徴量の値が高い傾向，負の値で

1）　sonoisa/sentence-bert-base-ja-mean-tokens-v2
2）　koheiduck/bert-japanese-finetuned-sentiment

表 5　分類における特徴量の重要度と効果量

	詩		歌詞		短編小説	
特徴量	重要度	効果量	重要度	効果量	重要度	効果量
Maas Index	0.049	**0.405**	**0.081**	**0.494**	0.038	-0.285
MTLD	0.049	**-0.384**	0.058	**-0.402**	0.029	-0.056
MATTR	**0.077**	**-0.440**	0.035	-0.297	0.027	0.004
HD-D	0.053	-0.042	0.039	-0.160	**0.114**	**0.597**
Hapax Legomenon Ratio	0.046	**-0.424**	**0.097**	**-0.538**	0.039	**-0.326**
Lexical Density	**0.090**	**-0.475**	0.033	0.029	0.043	-0.144
文長のばらつき	**0.167**	**0.577**	**0.252**	**0.695**	**0.357**	**0.854**
音素の美的スコア	0.037	-0.199	0.028	0.052	0.021	**0.305**
修飾関係の意味的距離の平均	0.042	-0.263	0.031	-0.016	0.028	0.205
修飾関係の意味的距離の分散	0.029	0.121	0.025	0.134	**0.097**	**0.548**
修飾関係の意味的距離の範囲	0.033	0.007	0.025	0.017	0.037	**0.476**
具象性スコア	0.035	-0.120	0.040	0.098	0.024	0.131
隣接文類似度の平均	0.032	0.210	0.028	0.227	**0.078**	**0.513**
隣接文類似度の分散	0.052	**0.357**	0.054	**0.357**	0.019	0.276
隣接文類似度の範囲	0.034	**0.310**	0.045	**0.363**	0.013	0.201
感情スコアの分散	**0.119**	**0.492**	**0.080**	**0.390**	0.015	0.121
感情スコアの範囲	0.042	**0.381**	0.031	0.248	0.011	-0.001
五感表現の割合	0.016	-0.051	0.017	-0.137	0.010	-0.039

あれば LLM の方が特徴量の値が高い傾向であることを示している．効果量の絶対値が 0.3 を超えている値を太字で示す．

全ての文学テキストの形式に共通する有効な特徴は文長のばらつきで，LLM テキストの方が文長のばらつきが低い，つまり，LLM は同じ長さの文を生成する傾向があると言える．それ以外の有効な特徴では，詩と歌詞でほぼ同じになっている．詩と歌詞に共通して有効な特徴は感情スコアの分散で，LLM テキストの方が感情の分散が低い，つまり，LLM の生成する詩・歌詞は文章全体を通じて感情の起伏が乏しい傾向がある．また，詩・歌詞のそれぞれで有効な語彙指標に注目すると，LLM が生成する詩・歌詞は，人間が作るものよりも語彙が多様である傾向があると言える．

一方で，短編小説は詩・歌詞と異なり，隣接文類似度の平均と修飾関係の意味的距離の分散が有効な特徴となっている．LLM の生成する短編小説は，人間の作るものよりも文章の首尾一貫性が低い傾向がある．また，修飾のパターンの振れ幅が少なく，表現の起伏が少ない傾向がある．そして，有効な語彙指標に注目すると，LLM が生成する短編小説は，人間が作るものよりも語彙が乏しい傾向があると言える．

6　おわりに

本研究では，人間が作成した文学テキストと LLM が生成した文学テキストの BERT による分類が高性能で可能であることを示し，人間と LLM が生成した文学テキストの間に存在する識別可能な特徴について分析した．LLM の生成する文学テキストは人間よりも文長が一定であり，詩と歌詞については，人間よりも語彙が多様だが，感情の変動が小さい傾向がある．また，LLM の生成する短編小説は，人間よりも文章の首尾一貫性が低い傾向があり，修飾の多様性や語彙の多様性に欠けている傾向があることが示された．

LLM が生成する文学テキストの傾向を読み取ることで，教育面において，学生が自分の力で文章を作成することを促し，思考力や創造性を守ることができると考えられる．また，文学的創造性やオリジナリティの所在を再考する手がかりを与える．

今後の課題としては，LLM の生成する文学テキストの特徴を回避するようなプロンプトで文学テキストを生成した場合，どのくらい人間の文学テキストに近づくのかということが挙げられる．

参考文献

[1] Brian Porter and Edouard Machery. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. **Scientific Reports**, Vol. 14, No. 1, p. 26133, 2024.

[2] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. A survey on LLM-generated text detection: Necessity, methods, and future directions. **Computational Linguistics**, Vol. 51, No. 1, pp. 275–338, 03 2025.

[3] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, and Toru Sasaki. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. **arXiv preprint arXiv:2305.14902**, 2023.

[4] Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. Mgtbench: Benchmarking machine-generated text detection. In **Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security**, CCS '24, p. 2251–2265, New York, NY, USA, 2024. Association for Computing Machinery.

[5] Afendi Hamat. The language of ai and human poetry: A comparative lexicometric study. **3L, Language, Linguistics, Literature**, Vol. 30, No. 2, pp. 1–20, 2024.

[6] Philip M McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. **Behavior research methods**, Vol. 42, No. 2, pp. 381–392, 2010.

[7] Theresa Matzinger and David Košić. Phonemic composition influences words' aesthetic appeal and memorability. **PLoS One**, Vol. 20, No. 12, p. e0336597, 2025.

[8] NAIST Social Computing Lab. 日本語抽象度辞書「AWD-J: Abstractness of Word Database for Japanese common words」, （2025-12 閲覧）. https://sociocom.naist.jp/awd-j/.

A　語彙指標の定義

Maas Index　語彙の複雑さの逆の尺度．値が高いほど，言語のシンプルさが増していることを示す．N を総単語数，V を異なり語数とすると次の式で表せる．

$$a^2 = \frac{\log_{10} N - \log_{10} V}{\log_{10}^2 N} \tag{1}$$

MATTR　テキスト長の問題に対処した Type-to-Token Ratio(TTR) の派生指標．w をウィンドウサイズ，V_i を i 番目のウィンドウ内の異なり語数，N を総単語数とすると次の式で表せる．

$$\mathrm{TTR} = \frac{V}{N} \tag{2}$$

$$\mathrm{MATTR} = \frac{1}{N-w+1} \sum_{i=1}^{N-w+1} \frac{V_i}{w} \tag{3}$$

MTLD　TTR が閾値（0.72）を下回るまでの長さを基準に計算する．テキストの先頭から TTR を計算する．TTR が 0.72 を下回った場合，そこまでを 1 つの要因としてカウントする．テキストの最後まで行い，要因数（FS）を求める．そして，テキストを逆順にして，同様に計算する．順方向と逆方向の平均値が MTLD となる．FS を算出された要因の総数，N を総単語数とすると次の式で表せる．

$$\mathrm{MTLD} = \frac{N}{\mathrm{FS}} \tag{4}$$

HD-D　語彙の頻度帯域全体にわたる単語の分布に関する指標．値が高いほど稀な単語の使用が多いことを示す．テキストからランダムに k 個の単語を抽出した際，各単語タイプ（異なり語）が少なくとも 1 回出現する確率を全語彙について合計したもの．テキスト内の全単語タイプの集合を V，テキスト全体における単語タイプ t の出現頻度を $f(t)$，サンプルサイズを k とすると次の式で表せる．

$$\text{HD-D} = \sum_{t \in V} \left(1 - \frac{\binom{N-f(t)}{k}}{\binom{N}{k}} \right) \tag{5}$$

Hapax Legomenon Ratio　テキストの全トークン数に対する一度だけ出現する固有の単語の比率を示す．値が高いほどより豊かな固有の単語の使用を示す．V_1 を一度しか出現しない単語数，N を総単語数とすると次の式で表せる．

$$\text{Hapax Legomenon Ratio} = \frac{V_1}{N} \tag{6}$$

Lexical Density　テキスト中の内容語（名詞，動詞，形容詞，副詞）の割合を示す．N_{content} を内容語の総数，N を総単語数とすると次の式で表せる．

$$\text{Lexical Density} = \frac{N_{\text{content}}}{N} \tag{7}$$

B　語彙指標以外の定義

文長のばらつき　文長の変化に関する指標．1 つの文ごとに文字数を数え，分散を求める．

音素の美的スコア　音素の美的魅力に関する指標．美的価値の高い音素を 1 点，低い音素を-1 点，中立の音素を 0 点とし，テキスト全体の総和を求める．そして，総和を音素単位の総数で割った平均を美的スコアとする．

修飾関係の意味的距離　分布意味論に基づいた指標．係り受け解析を行い，文法的に強い関係（主述関係，名詞修飾，形容詞修飾，連用修飾）のあるペアを抽出し，GiNZA の単語ベクトルを基にコサイン類似度を求める．そして，1 からコサイン類似度を引いた値の平均，分散，範囲（最大値と最小値の差）をそれぞれ求める．

具象性スコア　単語が具体的かどうかを評価する指標．まず，日本語抽象度辞書に掲載されている抽象度スコアを基に，辞書にヒットした単語の具象性スコアの合計を求める．なお，日本語抽象度辞書では，1 が最も具体的で，5 が最も抽象的と定義されているため，数値を反転させている．そして，その合計に対して辞書でヒットした単語の総数で割った値を具象性スコアとする．

隣接文の類似度　文脈変化に関する指標．隣り合う文 S_i と S_{i+1} のコサイン類似度を計算し，平均と分散と範囲を求める．

感情の揺らぎ　テキスト中の感情変化に関する指標．BERT ベースのモデルで，1 つの文ごとにポジティブ (+1)〜ネガティブ (−1) にスコアリングする．そして，それらの分散と範囲をそれぞれ求める．

五感表現の割合　五感に関する表現が使われているかどうかを評価する指標．日本語 WordNet を利用し，五感（視覚，聴覚，嗅覚，味覚，触覚）の各モダリティにおいて，知覚動作，感覚器，感覚特性に対応する最上位の単語を選定し，それらの関連語と合わせて五感語辞書を作成する．作成した辞書を用いて単語マッチングを行い，ヒットした回数を求める．そして，その回数を内容語（名詞，動詞，形容詞）の総数で割った値を五感表現の割合とする．