ISSN 2186-5647

2-6

中国語と日本語における大規模言語モデル生成文検出の比較

日大生産工(院)　○LI SHUYING　　日大生産工　栃窪 孝也

1. まえがき

近年、自然言語処理(NLP)の分野では大規模言語モデル(LLM)が飛躍的に発展し、GPT-3.5やGPT-4などに代表されるモデルは膨大なパラメータと学習データを基盤に、人間に匹敵する自然な文章生成を実現している[1]。その結果、対話システムや機械翻訳、自動要約や教育・研究支援など多様な応用が急速に拡大している。

一方で、この性能向上は新たな課題も生み出している。すなわち、LLMが生成したテキストが「人間の文章」か「人工的に生成された文章」かを判別することが難しくなっている。従来の生成文は不自然な表現や統計的偏りが目立ったが、最新のLLMは文脈理解や文体の一貫性が飛躍的に向上し、人間との差がほとんどない。このため、教育現場では評価の不正確化、研究分野では剽窃や捏造の助長、社会全体ではフェイクニュースや詐欺メールの大量生成が懸念されている[2]。

このような背景から、近年は「LLM生成文の検出技術」に関する研究が進展している。しかし、これらの研究の大半は英語を対象としており、日本語や中国語といった非英語言語の検証は限られている[3]。英語は語形変化が単純で単語境界が明確なため、統計的特徴量を直接適用しても高い性能を得やすい。一方、日本語や中国語は形態素解析や単語分割を要し、構造が大きく異なる。日本語は膠着語で助詞・助動詞により文脈が決まるため、単語頻度のみでは特徴を捉えにくい。中国語は単語境界が曖昧で文構造も異なることから、利用可能な特徴も異なる[4]。

このように、非英語言語でのLLM生成文検出は依然として未開拓である。特に日本語と中国語は文法体系が大きく異なるため、比較研究により言語的特徴が検出性能に与える影響を明らかにできると考えられる。そこで本研究では、日本語と中国語における大規模言語モデル生成文検出の比較を目的とし、RoBERTaと統計的特徴量に基づく手法を適用し、ROC-AUCを中心に性能を評価する。

2. 関連研究

大規模言語モデル(LLM)の普及に伴い、生成文検出の研究は近年急速に発展している。英語を中心に、多くの研究が「人間が書いた文」と「LLMが生成した文」を識別するための手法を提案してきた。これらの研究は大きく二つの系統に分類できる。

一つ目は電子透かし方式である。この手法では、LLMが文章を生成する際に、確率分布の制御やトークン選択の調整を通じて、人間には知覚できない「透かし」を文中に埋め込む。その後、専用の検出器を用いることで、その文章が特定のモデルによって生成されたか否かを判別できる。

この方式は理論的に高い検出性能を期待できる一方で、モデル内部の改変が必要であるため外部から利用できないこと、また既に生成・流通しているテキストには適用できないという制約が存在する。

二つ目は特徴量分析および分類モデル方式である。この手法では文章そのものを解析対象とし、統計的特徴や深層学習モデルを用いて識別を行う。代表的な統計的特徴には、対数尤度(Log-Likelihood)、生成確率順位(Rank)、エントロピー(Entropy)などがある。

例えば、Gehrmannらは、文章中の各単語が言語モデルにおいて高頻度で生成されるか否かを可視化し、人間による検出を支援する仕組みを提案した[2]。また、Mitchellらは、生成文に微小な摂動を加えた際の確率曲線の変化を利用することで、ゼロショットでも高い識別性能を達成している[1]。

日本語を対象とした研究としては、丸井らが「日本語における大規模言語モデルの生成文検出」を発表している[3]。この研究では、「Yahoo!知恵袋データセット」を用いて人間文7500件と大規模言語モデル(GPT-3.5-turbo)が生成した文7500件を組み合わせたデータセットを構築している。RoBERTaモデルと統計的特徴量に基づく手法を評価した結果、RoBERTaはROC-AUCが0.998と最も高い性能を示し、統計的手法も一定の有効性を示した。

深層学習を用いた生成文検出では、BERTが

A Comparative Study on Detecting Machine-Generated Text in Chinese and Japanese Using Large Language Models

Shuying LI and Kouya TOCHIKUBO

質問	人の答え	AIの答え
SNSについての質問です YouTubeをする人はユーチューバー、Instagramをする人はインスタグラマー、TikTokをする人はティックトッカーと呼ばれますが、Twitter(X)やビーリアルをする人はなにか呼ばれ方があるのでしょうか？	"Twitter(X)はツイッタラーという呼ばれ方が一般的かと思われます。 BeRealのほうはビーリアラーと呼ぶ人もいるようですが、あまり見かけない印象です。"	"Twitterをする人はツイッタラー、ビーリアルをする人はビーリアラーと呼ばれることがあります。ただし、それぞれのSNSに特有の呼び方は確立されていない場合もあります。"

表１　日本語データの例

基盤となっている。BERTは文脈の双方向情報を同時に学習できる点に特徴があり、自然言語理解タスクにおいて高い性能を発揮している。これを改良したRoBERTaは、学習データ量と学習時間を増やし、次文予測のタスクを削除するなどの最適化によって、より高い表現能力を実現している。

既存研究[3]では、日本語を対象としてRoBERTaおよび統計的特徴量に基づく手法の性能を比較し、その有効性を確認している。しかし、この研究は単一言語(日本語)に限定されており、言語構造の異なる多言語への適用可能性については検討されていない。

一方、本研究では、既存手法を基盤としつつ、日本語と中国語を比較対象とする多言語的検証を行う点に特徴がある。特に、語順や形態構造の異なる両言語における識別性能の差を分析することで、言語構造が検出特性に及ぼす影響を明らかにすることを目的とする。

3.　データセット

本研究で使用したデータセットは、質問と回答から構成されるQ&A形式を採用した。

日本語データについては、「Yahoo!知恵袋」から人間による回答文を500件収集し、それぞれの質問文を大規模言語モデルGPT-3.5-turboに入力して生成文500件を作成した。日本語データの例を表１に示す。

一方、中国語データについては、中国のQ&Aプラットフォーム「知乎(Zhihu)」から人間回答500件を収集し、日本語データと同様の手順により、対応する質問文をGPT-3.5-turboに入力して生成文500件を作成した。これにより、両言語間で同一形式・同一規模のデータ構造を保持した比較が可能となった。

以上の手順により、各言語につき、人間文500件と生成文500件からなる計1000件のデータセットを構築した。

4.　検証手法と実験方法

4.1　RoBERTaを用いた深層学習モデル

本研究では、日本語データに対しては早稲田大学河原研究室が公開している「nlp-waseda/roberta-base-japanese」を用い、中国語データに対しては哈工大・訊飛聯合実験室(HFL, iFLYTEK)によって公開されている「hfl/chinese-roberta-wwm-ext」を利用した。

4.2　統計的特徴量を用いた手法

RoBERTaのような深層学習モデルは高精度である一方、大量の計算資源を必要とするという課題がある。そこで本研究では、事前学習を必要とせず、言語モデルの確率分布に基づいて計算される統計的特徴量を利用する手法も検証対象とした。この方法は、生成文は人間文に比べて「高確率の語彙を選択しやすい」という仮定に基づいている。

具体的には、文をトークン列$\omega_1, \omega_2, \cdots, \omega_T$としたとき、以下の式で計算される[3]。

(1) 対数尤度

$$\frac{1}{T}\sum_{i=1}^{T}\log p(\omega_i|\omega_{1:i-1}) \tag{1}$$

(2) 生成確率の順位

$$-\frac{1}{T}\sum_{i=1}^{T}\sum_{\omega\in V} I\left(p(\omega|\omega_{1:i-1}) < p(\omega_i|\omega_{1:i-1})\right) \tag{2}$$

(3) 生成確率の対数順位

$$-\frac{1}{T}\sum_{i=1}^{T}\log\left[\sum_{\omega\in V} I\left\{p(\omega|\omega_{1:i-1}) < p(\omega_i|\omega_{1:i-1})\right\}+1\right] \tag{3}$$

(4) 生成分布のエントロピー

$$\frac{1}{T}\sum_{i=1}^{T}\sum_{\omega\in V} p(\omega|\omega_{1:i-1})\log p(\omega|\omega_{1:i-1}) \tag{4}$$

5.　実験方法

実験に用いたデータセットは、構築した日本語および中国語のQ&Aコーパスを基盤としたものである。

RoBERTa系モデルを用い、それぞれ本研究のデータセットに対してファインチューニングを実施した。学習に用いたハイパーパラメータを表２に示す。

各文に対して、対数尤度、生成確率順位、生成確率の対数順位、エントロピーの四種類の特徴量を計算した。その後、閾値を段階的に変化させてROC曲線を描き、AUCを算出することで検出性能を評価した。

ROC曲線は、分類の閾値を変化させたときの真陽性率(TPR)と偽陽性率(FPR)の関係を表すものであり、AUC値はその曲線下の面積を示す。AUCが0.5であればランダム分類と同等であり、1.0に近づくほど高い検出性能を示す。

表２　学習時のハイパーパラメータ

項目	値
学習率	5e-5
エポック数	3
バッチサイズ	8
ドロップアウト率	0.1
最適化手法	AdamW
損失関数	二値クロスエントロピー

6.　実験結果

RoBERTaモデルおよび統計的特徴量に基づく手法を用い、日本語および中国語それぞれ1,000件(人間文500件＋生成文500件)のデータを対象に検出実験を行った。性能評価にはROC-AUCを用いた。

RoBERTaモデルは日本語で0.926、中国語で0.999と高い識別性能を示した。統計的手法においても、中国語はすべての指標で日本語を上回った。さらに既存研究(7,500件＋7,500件)[3]との比較により、本研究の中規模データ実験でも高い有効性が確認された。結果を表３に示す。

各言語におけるROC曲線を図１および図２に示す。

図１は日本語データに対する結果であり、RoBERTaモデルのAUCは0.926を示した。曲線は左上方向に張り付き、全体として高い識別性能を示しているが、一部閾値領域では偽陽性率(FPR)がやや上昇しており、文体的曖昧性を反映していると考えられる。

一方、図２の中国語データでは、全モデルにおいてROC曲線がほぼ左上に一致しており、AUCが0.999と極めて高い値を示した。特に統計的特徴量(対数尤度・生成確率順位など)でも高精度を維持しており、文構造が単純で語順依存性が強い中国語の特徴を反映している。

これらの結果は、言語の構造的違いが検出特性に直接影響を及ぼすことを示唆している。

表３　日本語中国語比較結果

	日本語	中国語	既存研究
データ数	500+500	500+500	7500+7500
RoBERTa	0.926	0.999	0.998
対数尤度	0.915	0.975	0.847
生成確率順位	0.740	0.972	0.793
生成確率対数順位	0.899	0.976	0.890
エントロピー	0.876	0.934	0.811

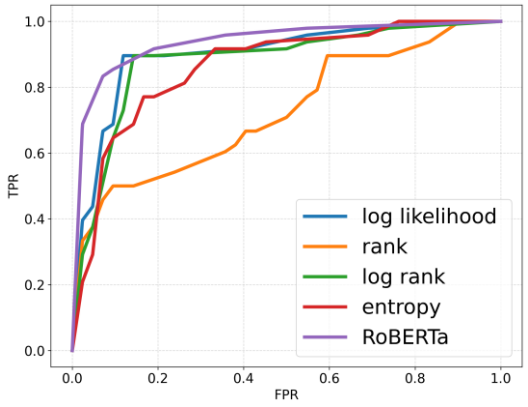

図１　日本語データにおける ROC 曲線

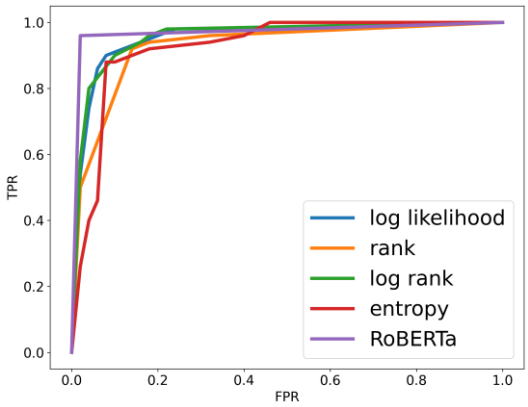

図２　中国語データにおける ROC 曲線

7. 日本語と中国語の結果比較

表２に示すように、中国語はすべての評価指標で日本語を上回り、統計的特徴量の有効性が特に顕著であった。以下では、この性能差の要因を言語的側面から考察する。

(1) 言語構造の違い

中国語は主語–述語–目的語(SVO)の語順を持ち、文構造が比較的単純で、語順に従った統計的特徴が強く表れる傾向がある。そのため、生成文が出現確率の高い語彙に偏る特徴が顕著に現れやすく、統計的指標が有効に機能する。

日本語は膠着語であり、助詞や助動詞の多様な組み合わせによって意味関係が構築される。このため、語順の自由度が高く、統計的指標では人間文と生成文の差が曖昧になりやすい。

(2) 語境界の扱い

本研究では分かち書きツールを用いずに実験を行ったため、言語固有のトークン化処理の違いが影響した。中国語は1文字ごとに意味を持つ場合が多く、文字単位の確率分布が安定しやすい。対照的に日本語は漢字・ひらがな・カタカナが混在し、文字単位の分布が不安定で、統計的特徴の精度が低下した可能性がある。

(3) 文末表現の多様性

日本語には「〜と思う」「〜である」「〜でしょう」など多様な文末表現が存在し、意味的に同じでも異なる表現が多数用いられる。これにより、生成文と人間文の違いが統計的に曖昧になりやすい。一方、中国語では文末表現の種類が比較的限定されるため、確率的偏りが顕著に現れたと考えられる。

8. まとめ

本研究では、日本語と中国語における大規模言語モデル(LLM) 生成文検出の比較を行い、RoBERTaを用いた深層学習モデルと統計的特徴量に基づく手法を対象に性能評価を行った。人間文500件と生成文500件を用いた実験の結果は以下の通りである。

RoBERTaは両言語において高い検出性能を示し、とりわけ中国語ではROC-AUCが0.998とほぼ完全に近い識別が可能であった。日本語でも0.926と高い値を記録し、深層学習モデルが生成文検出に有効であることが改めて確認された。

統計的特徴量に基づく手法の有効性は言語によって差が見られた。中国語では対数尤度や生成確率順位といった特徴量が0.90前後の高いAUCを示した一方、日本語では0.80前後に留まり、特に生成確率順位では0.715と精度が低かった。この結果は、言語構造の違いや文末表現の多様性、表記体系の複雑さなどにより、日本語においては単純な確率的指標では十分な識別が困難であることを示している。

既存研究の大規模データ実験と比較した結果、RoBERTaはデータ規模に依存せず安定して高性能を維持できることが確認された。一方、統計的手法は大規模データでは精度が低下する傾向が見られ、中規模データ条件でより有効に機能する可能性が示唆された。

これらの結果から、本研究は以下の意義を持つ。すなわち、非英語言語における生成文検出の比較を通じて、言語特性が検出性能に及ぼす影響を実証的に明らかにした点である。特に、日本語と中国語は文法体系が大きく異なるため、両言語を対照することで検出手法の普遍性と限界を浮き彫りにすることができた。

今後の課題としてはより多様な生成モデル(GPT-4、LLaMA、Claudeなど)を対象とした評価を行い、検出手法の一般性を検証する必要がある。また、敵対的プロンプトやリライトなど、生成文のスタイルを意図的に操作したケースへの耐性を評価することが重要である。さらに、日中以外の多言語比較を進めることで、多言語対応型の検出技術の設計指針を確立できると期待される。

以上の成果は、日本語と中国語における生成文検出研究の基礎的な成果を示すものであり、今後の多言語的かつ実運用的な検出技術の発展に寄与するものである。

参考文献

1) Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, Chelsea Finn, “DetectGPT: Zero-shot machine-generated text detection using probability curvature,” Proc, 40th. International Conference on Machine Learning, (2023) pp.24950-24962.
2) Sebastian Gehrmann, Hendrik Strobelt, Alexander Rush, “GLTR: Statistical detection and visualization of generated text,” Proc, 57th. Annual Meeting of the Association for Computational Linguistics: System Demonstrations, (2019) pp.111-116.
3) 丸井渚生, 曹洋, 中村篤祥, “日本語における大規模言語モデルの生成文検出,” DEIM, No.16, (2024) pp. 718-722.
4) Emily M. Bender, “On Achieving and Evaluating Language-Independence in NLP,” Linguistic Issues in Language Technology, Vol.6, (2011) pp.1-28.