TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　自然言語の非線形性の計算論モデル
英文：Computational modeling of Natural Language Nonlinearity

利用課題責任者　田中久美子
Kumiko TANAKA-ISHII

早稲田大学　理工学術院　基幹理工学部
Waseda University, School of Fundamental Science and Engineering
URL　https://www.ml-waseda.jp/

邦文抄録（300 字程度）

本研究は、自然言語の非線形性を数理的に解析し、計算論的モデルとして表現することを目的とする。大規模言語モデル(LLM)の発展に伴い、自然言語の複雑性を精緻に定量化する手法の重要性が高まっている。本研究では、言語の自己相似性に着目し、相関次元を測定することで言語の複雑性を評価する手法を提案した。日本語・英語・ドイツ語・中国語の四言語において相関次元が約 6.5 に収束することを確認し、異なる言語間での共通する非線形な特性の存在を示唆した。さらに、生成モデルを活用した文書分類・クラスタリング手法を開発し、言語の複雑性を考慮した新たな情報処理技術の可能性を示した。今後の課題として、相関次元の形成要因をより詳細に解明し、言語モデルの推論能力を定量化する指標の開発を進める。

英文抄録（100 words 程度）

This study mathematically analyzes the nonlinearity of natural language and develops computational models to represent its complexity. With the advancement of large language models (LLMs), precise quantification of language complexity has become crucial. We propose a method that evaluates textual complexity through correlation dimension, capturing self-similarity in natural language. Our findings indicate a stable correlation dimension of approximately 6.5 across Japanese, English, German, and Chinese, suggesting a universal nonlinear characteristic. Additionally, we develop a document classification and clustering approach leveraging LLMs. Future work will focus on elucidating the formation mechanisms of correlation dimensions and establishing metrics to quantify reasoning capabilities.

Keywords: 自然言語処理, 生成モデル, 複雑性, 非線形性, 相関次元.

背景と目的

自然言語は高度に非線形な構造を持ち、その複雑性は計算的に解析しづらい特性を有する。この非線形性は、単語頻度の処理だけでは本質的な構造を理解することはできない。近年、大規模言語モデル(LLM)の発展により、言語の複雑性を数学的に解析する新たな可能性が開かれている。これにより、従来の言語モデリング手法では見えなかった構造が捉えられるようになり、言語の階層的特性や自己相似性といった要素の理解が進んでいる。

本研究では、自然言語の本質的な構造を数理的に定式化し、計算モデルを構築することで、言語の特性を精緻に捉えることを目的とする。また、本手法を法律・金融分野の応用に展開し、実際の情報処理技術に組み込むことで、言語理解・検索・知識抽出の効率向上を目指し、実用的なモデルの構築を行う。

概要

深層学習技術の進展により、大規模言語モデル（LLM）が提案され、情報科学全域に新たな可能性をもたらしている。LLM の性能が向上するにつれて、その限界や課題も明確になりつつある。その一つに、LLM が言語の「非線形」な特質をどの程度適切に学習・再現できるかという問題がある。従来の自己回帰型の言語モデルは、テキストの局所的な確率分布を学習することには長けているが、グローバルな言語構造を適切に捉えることが難しい。

本研究では、言語の「非線形性」に着目し、数学的・計算的手法を用いた解析を行う。特に、相関次元を用いた言語の複雑性定量化手法を開発し、異なる言語間での比較を試みる。このアプローチは、単なる情報エン

トロピーや perplexity（困惑度）とは異なり、言語が持つ階層的構造や自己相似性を測定することが可能である。本研究では、この手法を大規模言語モデルに適用し、LLM が学習する言語の構造がどのように進化するかを詳細に解析する。

また、生成モデルを活用した文書クラスタリング・検索手法を開発し、言語の複雑性を考慮した新たな情報処理手法の可能性を探る。これにより、従来のベクトル空間モデルに基づく手法を超え、生成モデルの持つ強力な表現力を活かした新たな検索・クラスタリング手法の実現を目指す。

結果および考察

本研究では、相関次元を用いた言語の複雑性定量化手法を提案し、日本語・英語・ドイツ語・中国語の四言語について分析を行った。その結果、各言語の相関次元が約 6.5 に収束すること(図 1)を確認し、異なる言語間で共通する非線形な特性が存在する可能性を示唆した。また、相関次元と長期記憶性の関連性を示し、言語の持つ情報構造が、単なる統計的特徴だけでなく、より深いレベルでの記憶性を反映していることを明らかにした。本研究の成果は Physical Review Research に掲載され[1]、さらにフラクタル研究の国際的なワークショップ Geometry and Stochastics にて招待講演[2]として発表された。

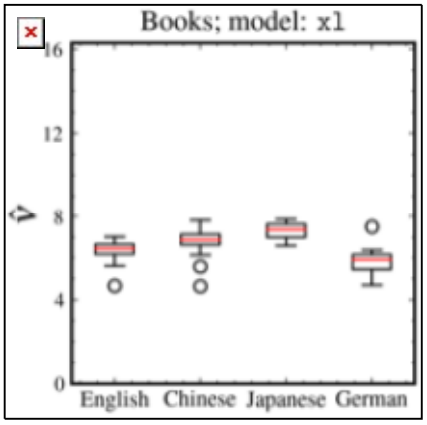

図 1 四言語テキストの相関次元

また、生成モデルを用いた文書クラスタリング・検索手法を開発し、言語の複雑性を考慮した新たな情報処理手法の可能性を探った。特に、新たなクラスタリング手法として「生成クラスタリング」を提案した。従来のクラスタリング手法では、主にベクトル埋め込みを用いた手法が一般的であったが、本手法では生成モデルを活用し、情報論的なアプローチに基づく新たな文書表現方法を提案した。その結果、従来の埋め込み方法と比較して大幅な改善が見られ(図 2)、情報検索や知識抽出における応用可能性が大きく広がった。本研究の成果は、世界トップの機械学習・人工知能学会である ICML[3]および AAAI[4]に掲載された。

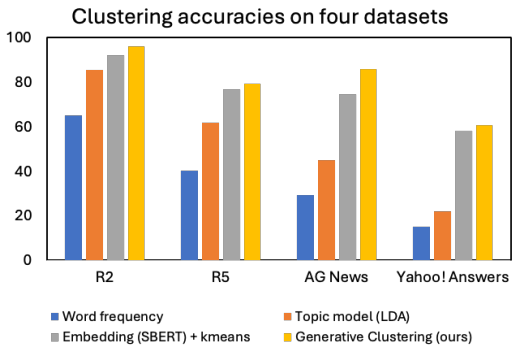

図 2 四データセット上で各クラスタリング方法の精度

まとめ、今後の課題

本研究では、相関次元を用いた言語の複雑性定量化手法を開発し、その有効性を大規模データセットに適用することで検証した。また、生成モデルを活用した文書クラスタリング・検索手法を開発し、新たな情報処理手法の可能性を示した。

今後の課題として、相関次元の形成要因の解明を進め、言語の階層的構造や自己相似性との関連を明確にする必要がある。また、相関次元を LLM の推論能力の評価指標として確立し、生成クラスタリングの改良を通じて法律・金融分野への応用を実用化することが求められる。さらに、LLM の学習ダイナミクスを非線形性の観点から分析し、モデルの特性や限界をより深く理解することで、自然言語処理技術の発展に寄与することを目指す。

参考文献

[1] Xin Du and Kumiko Tanaka-Ishii. Correlation dimension of natural language in a statistical manifold. Physical Reviews Research, 6 (2), L022028.
[2] Kumiko Tanaka-Ishii and Xin Du. Correlation dimension of large language models. In Proceedings of Fractal Geometry and Stochastics 7, 2024. Invited talk.
[3] Xin Du, Lixin Xiu, and Kumiko Tanaka-Ishii. Bottleneck-Minimal Indexing for Generative Document Retrieval. In Proceedings of ICML 2024.
[4] Xin Du and Kumiko Tanaka-Ishii. Information-Theoretic Generative Clustering of Documents. In Proceedings of AAAI 2025.