大規模言語モデルによるテキスト内データの関係グラフ構築

1255068　鈴木 亮佑

Relation-graph construction of text data using a large language model

Ryosuke Suzuki

【背景】2017 年に深層学習モデル「Transformer」が新たに発表され,翌年 2018 年には Transformer を使用した事前学習モデル「BERT」が登場し,自然言語処理分野はブレイクスルーを迎えた.[1] その後,学習データと Transformer のパラメータ数の大規模化が進展し,高精度な大規模言語モデルが数多く登場した.その結果,大規模言語モデルは,文章要約,翻訳,あるいはテキストからの画像生成等の様々な下流タスクの基盤モデルとして利活用されている.

【目的】本研究では,モデルの学習過程で生成される Transformer のエンコードデータを用いた,単語間の関係グラフ（単語ネットワーク）を構築し有用情報を得ることを目指した.

【方法】テキストデータ例として,ginger の代表的化合物「gingerol」に関する技術文献群を使用した.Transformer より抽出したトークンのエンコードデータから,単語単位のエンコードデータを得るオリジナル手法を開発した.その上で,得られたエンコードデータより単語ペアの cos 類似度を算出し,重み付き単語ネットワークを得た.更に,辺の重みに関するハイパスフィルタリングにより,単語ペアの関係データに関するランク付けを実現した.

【結果】単語ネットワークのうち,gingerol ノードと近接した部分ネットワークに着目した.なお,単語ネットワークを構築する際,テキストデータに含まれる化合物リスト（技術文献で頻出）の影響を軽減するため,共起頻度が 10 回未満の単語ペアを排除した.また,辺の重み（cos 類似度）として閾値 0.8 を採用することにより,ハイパスフィルタリングを行った.図 1 に,得られた単語ネットワークにおける,gingerol ノード周りの部分ネットワークを示す.これより,gingerol との関連度が高い化合物として,ginger の産業利用に関わる化合物群が抽出されることが確認された.また図 2 は,当該ネットワークの次数分布に関する対数プロットである.図 2 より,得られた単語ネットワークは高次数ハブノードを有しており,スケールフリー的性質を保持していることが見て取れる.

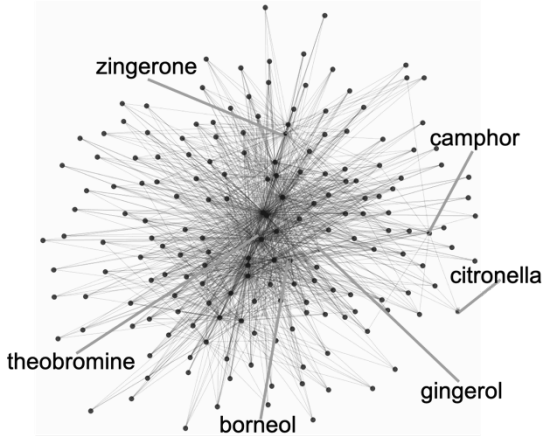

図 1 gingerol 周りの部分グラフ

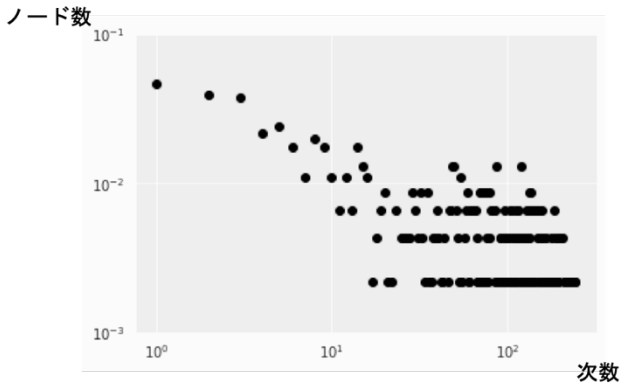

図 2 次数分布

文献

1) Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).