日本語における Transformer の認知的妥当性

Cognitive Validity of Transformer in Japanese

新海 功貴
指導教員　菊池 眞之

東京工科大学大学院　バイオ・情報メディア研究科
コンピュータサイエンス専攻　ブレインコンピューティング研究室

キーワード：脳波，EEG，事象関連電位，N400，Transformer，認知科学

１．はじめに

近年，ChatGPT を始めとする大規模言語処理モデルの台頭が著しいが，その基礎となったのは 2017 年に発表された Transformer である．それまでの言語モデルは，RNN(Recurrent Neural Network)をベースにモデル構築を行うことが主流だったが，並列処理との相性が悪いことがネックになっていた．そこで Vaswani ら[1]は，これまで RNN の中で限定的にしか採用されていなかった Attention をモデルのメイン機構とした Transformer を提案し，機械翻訳タスクにて SoTA を叩き出すと同時に，計算負荷コストも RNN より抑えることに成功した．

一方で，Transformer の Attention には，実際の人の言語処理とは乖離があるという指摘がされている．RNN は過去単語の情報をワーキングメモリに収まる範囲に圧縮しながら，次単語予測を行うのに対して，Transformer の Attention は一文中の過去単語の全情報を用いて次単語予測を行う機構になっており，人間のワーキングメモリの容量を考慮すると，情報が多すぎるという懸念がある．

言語モデルと人間の言語処理の間に，どのような関係性があるのかを検証する方法として，脳波の N400 と言語モデルの Surprisal との関係を分析する方法がある．N400 は，EEG で計測される事象関連電位の一種であり，文脈上で予測できなかった単語を読んだ 400 ミリ秒後に発生する負の電位である．この負の電位は，より予測が困難な単語であるほど振幅が大きく観測される特徴がある．情報量 Surprisal は言語モデルの次単語予測の困難さを表すことができる．Surprisal の算出式を以下に示す：

$$\mathrm{Surprisal} = -\log P(W_i \mid W_1 \ldots W_{i-1}) .$$

W は連続的な単語列を表している．i 番目の単語を予測する際に，先頭の単語から i – 1 番目の単語までを判断材料に予測確率を算出し，その値に負の対数をとったものが Surprisal である．

Merkx ら[2]は，英語における N400 と Surprisal の関係を対数尤度比で分析し，Transformer と GRU(RNN)で比較した結果，Transformer の方が N400 への説明能力が高いことを示した．しかし，これは言語の異なる日本語で同じように適用されるとは限らず，栗林ら[3]の研究では，英語では見られなかった日本語における Surprisal と人の読み時間に乖離があることが示された．

本研究では，日本語における N400 が言語モデルの Surprisal とどのような関係にあるのかを，実際に脳波計測を行い分析する．

２．実験

２.1．脳波計測

日本語の N400 計測は，Frank ら[4]の取り組みをベースに行う． 被験者数は田中ら[5]を参考に 5 人以上起用する．本実験で使用する脳波計は，ヘッドセットには Ultracortex Mark IV EEG Headset，アン

プには Cyton + Daisy Biosensing Boards を使用し，電極には ThinkPulse Active Electrode を用いる．この電極を 10-20 法に則った頭頂部付近の 5 電極（Fz，Cz，Pz，C3，C4）に配置し，計測を行う．

２．２．被験者のタスク

脳波計測をする際の呈示刺激を図 1 に示す．被験者にはモニターに表示された文章を読んでいくように指示し，500 ミリ秒ごとに一単語ずつ表示していく．被験者にはノイズの混入を防ぐため，実験中は瞬きなどの動作をできるだけ抑えるように指示をするが，一文の表示が終了する度に，画面中央に“＋”を表示し，瞬きなどの動作を行っても良い休憩時間を設ける．被験者が意味的不一致語を感じた際には N400 が計測される．

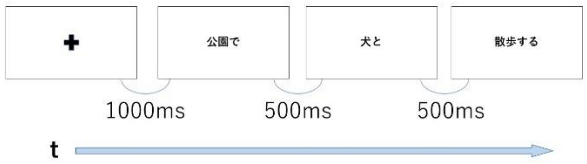

図 1 刺激呈示の例

２．３．言語モデルの準備

言語モデルについては，栗林ら[3]を参考に PyTorch 上で Transformer と LSTM の学習を行う．Surprisal の計算を行うために，両モデルともに left-to-right で構築を行い，学習データには脳波計測の刺激呈示で使う単語を含んだ日本語 Wikipedia コーパスを利用する．

３．解析手法

脳波成分の N400 と言語モデルの Surprisal の関係を分析するために，N400 を目的変数とした一般化線形混合モデルを作成する．このモデルに変量効果として，Surprisal を加えた際の対数尤度比を解析の対象とする．この対数尤度比が高ければ，日本語において，言語モデルが脳波成分を説明可能で，認知的な妥当性が高いと判断できる．

４．研究の現状と今後の予定

現状の進捗は，意味的不一致を含んだ日本語文 25 文と正文の 25 文のデータセットを作成し，脳波計測を行うための刺激呈示プログラムの最終調整段階である．脳波計測については，実験で使う脳波計のセットアップが問題なく完了すれば，実際に被験者を集めて実験を行う． 言語モデルの準備については，PyTorch 上で LSTM と Transformer が問題なく動作することを確認している．今後は，刺激呈示単語を含んだ日本語 Wikipedia コーパスを利用し，それぞれのモデルで学習を行い，十分な精度が得られることを確認する．解析手法については，Merkx ら[2]の一般化線形混合モデルを参考に R 言語にてモデルの実装を計画している．

参考文献

[1] Ashish Vaswani et al., “Attention Is All You Need”, 31st Conference on Neural Information Processing Systems, 2017.

[2] Danny Merkx, Stefan L. Frank, “Human Sentence Processing: Recurrence or Attention?”, Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 12–22, 2021.

[3] 栗林樹生，大関洋平，伊藤拓海，“予測の正確な言語モデルがヒトらしいとは限らない”，言語処理学会第 27 回年次大会，2021．

[4] Stefan L. Frank et al., “The ERP Response to the Amount of Information Conveyed by Words in Sentences”, Brain and Language volume 140, pages 1-11，2015.

[5] 田中久弥，宮本一郎，長嶋祐二，“事象関連電位 N400 計測に基づく日本手話理解における意味処理分析”，計測自動制御学会論文集 Vol.44，pages 768-775，2008．