jh251007

単語間に区切りのない書写言語における係り受け解析エンジンの開発

安岡孝一（京都大学人文科学研究所附属人文情報学創新センター）

概要

BERT・RoBERTa・DeBERTa・GPT2・LLaMA などの言語モデルにおいては、テキストをトークンに区切って学習をおこなう必要があり、欧米諸語においては、空白で区切られた単語をトークンとみなすようなトークナイザが用いられる。しかし、日本語・中国語・タイ語など、単語の間に区切りがない書写言語においては、空白によるトークナイザを用いることができない。

本研究では、単語の間に区切りのない書写言語に対し、係り受け解析エンジンの解析精度を指標として、各言語に対するトークナイザと、それを用いた言語モデルの開発をおこなっている。日本語 ModernBERT モデルにおいては、ひらがなを単文字トークナイズする手法により、品詞付与・係り受け解析の精度が向上することがわかった。また、トークナイザを改造する手法は、単語の切れ目が空白とズレているような言語においても有効であり、英語 ModernBERT やポルトガル語 ModernBERT への適用もおこなった。

1　共同研究に関する情報

1.1　共同研究を実施した拠点名

- mdx I

1.2　課題分野

- データ科学・データ利活用課題分野

1.3　参加研究者の役割分担

安岡孝一：研究統括
山崎直樹：文法構築
二階堂善弘：コーパス校訂
師茂樹：デジタル処理
Christian Wittern：コーパス校訂
池田巧：文法構築
守岡知彦：デジタル処理
鈴木慎吾：コーパス校訂
李媛：コーパス校訂
劉冠偉：コーパス校訂

2　研究の目的と意義

日本語・中国語・タイ語など、単語の間に区切りのない書写言語に対し、形態素解析 (単語切りと品詞付与) および係り受け解析をおこなうシステムを開発する。

現代の自然言語処理においては、巨大なテキストコーパスをもとに BERT・RoBERTa・DeBERTa・GPT・LLaMA・Qwen などの言語モデルを学習させる、という手法が、解析精度の向上に寄与する。これらの言語モデルにおいては、テキストを「トークン」に区切って学習をおこなう必要があり、欧米諸語においては、単語を「トークン」とみなして区切るような

トークナイザが用いられる。これは、単語の間に空白があるような欧米諸語においては、ある意味、自然な手法だと考えられる。しかし、日本語・中国語・タイ語など、単語の間に区切りがない書写言語においては、空白によるトークナイザを用いることができない。

われわれの研究グループは、古典中国語 (漢文)・近代日本語・タイ語などの多言語文法解析に挑戦してきた。これまでの研究の結果、古典中国語に対しては、漢字 1 文字 1 文字を「トークン」とみなすようなトークナイザが、文法解析においても有効に機能することが明らかとなった。近代日本語に対しては、字種 (漢字・ひらがな・カタカナ) によって「トークン」幅を変える必要がある、ということまでは明らかになってきているものの、どのようなトークナイザが最適なのかは、まだ不明である。タイ語に対しては、タイ文字クラスター (**คลัสเตอร์อักษรไทย**) を基本単位として、その組合せを音節へと拡張する形で「トークン」を設計する手法が良さそうだが、それが文法解析に最適なトークナイザなのかどうかは、まだまだ研究が必要である。

3 当拠点公募型研究として実施した意義

本研究は、日本語・中国語・タイ語などの巨大なテキストコーパスに対し、GPU を長時間稼働して言語モデルの学習をおこなう必要がある。本拠点の mdx I は、GPU を 24 時間 365 日稼働し続けることのできる環境であり、本研究を飛躍的に進めることが可能となっている。

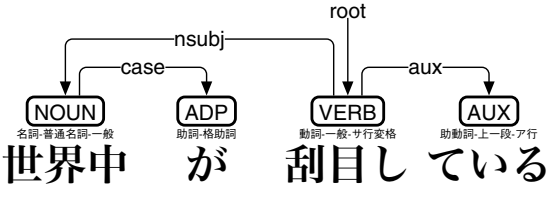
4 前年度までに得られた研究成果の概要

これまでわれわれは、BERT 系モデル (RoBERTa・DeBERTa) に対し、Biaffine などの隣接行列型アルゴリズムを用いて、係り受け解析エンジンを開発してきた。これに加え 2024 年度は、GPT 系モデル (GPT-2・GPT-Neo・GPT-NeoX・LLaMA・Qwen2) による係り受け解析エンジンの開発をおこなった。ただ、GPT 系モデルには、Biaffine はそのままの形では載せることができず、隣接確率行列を上三角行列に変換した形でのアルゴリズムを、新たに開発することになった。われわれとしては苦肉の策だったのだが、この新しい解析アルゴリズムは思いのほか解析精度が高く、しかも結果的に演算量を半分に減らすことができた。

一方、2024 年 12 月には、Answer.AI から ModernBERT が発表された。ModernBERT は BERT 系モデルの一種ではあるものの、内部的には GPT 系モデルの技術を大量に取り込んでいる。ならば、われわれの新しい解析アルゴリズムを、ModernBERT を含めた BERT 系モデルに適用し、それに合わせた「トークン」長を探りたい。これを 2025 年度のわれわれの目標とした。

5 今年度の研究成果の詳細

われわれの新しい解析アルゴリズムでは、品詞付与と係り受け解析を同時におこなう。単語間の各リンクに対する隣接行列に、品詞と係り受けラベルの両方を埋め込む形で解析をおこなう。ただし、隣接行列をそのまま用いるのではなく、上三角行列へと変換した上で、品詞付与と係り受け解析を同時におこなう。たとえば



という国語研長単位 Universal Dependencies (UD) 有向グラフに対する 4×4 の隣接行列

$$\begin{bmatrix} - & \underset{\text{case}}{\text{ADP}} & - & - \\ - & - & - & - \\ \underset{\text{nsubj}}{\text{NOUN}} & - & \underset{\text{root}}{\text{VERB}} & \underset{\text{aux}}{\text{AUX}} \\ - & - & - & - \end{bmatrix}$$

に対し、係り受けラベルにリンクの方向を付加した上で、左向きリンクの各要素を転置する。

$$\begin{bmatrix} - & \underset{\text{case}\rightarrow}{\text{ADP}} & \underset{\text{nsubj}\leftarrow}{\text{NOUN}} & - \\ - & - & - & - \\ - & - & \underset{\text{root}}{\text{VERB}} & \underset{\text{aux}\rightarrow}{\text{AUX}} \\ - & - & - & - \end{bmatrix}$$

この上三角行列の各行を一次元に展開し、系列ラベリングモデル上に実装する (図 1)。なお、入力側では、各行の末尾に [SEP] トークンを挟みこんでおく。

この品詞付与・係り受け解析アルゴリズムでは、UD 有向グラフのノード数 n に対し、入出力幅 $(n+1)(n+2)/2$ トークンの系列ラベリングモデルが必要[*1]となる。なお、実際の解析においては、空行は削除可能なことから、もう少し入出力幅を小さくできる。

この品詞付与・係り受け解析アルゴリズムを、各言語 ModernBERT に適用した。そうしたところ、解析精度の向上には、各言語ごとにトークナイザを改造する必要が生じた。

SB Intuitions 日本語 ModernBERT[*2]のトークナイザは、Sarashina シリーズで開発[*3]されたトークナイザを流用しており、トークン長が生成 AI 向きで、国語研長単位に合致していない。たとえば「世界中が刮目している」という文を「世界中」「が」「e5」「88」「ae」「目」「している」とトークナイズしてしまうため、「刮目し」「ている」の部分がうまくいかない。

図 1　上三角行列を用いた系列ラベリング

[*1] 入出力幅 8192 トークンの ModernBertForTokenClassification であれば、$n \leq 126$ の上三角行列を乗せることができる。

[*2] 塚越駿, 李聖哲, 福地成彦, 柴田知秀: 日本語 ModernBERT の構築, JLR2025『日本語言語資源の構築と利用』(2025 年 3 月 14 日).

[*3] https://www.sbintuitions.co.jp/blog/entry/2024/06/26/115641

国語研長単位における語境界をまたいでしまうと、どうしても解析精度が下がってしまう。

語境界をまたいでいる箇所を調べてみたところ、ほぼ全てがひらがなの前後だったので、ひらがなを単文字トークナイズすることにした。ただ、これだと今度はトークンが短くなりすぎるので、いったん仮の品詞付与を UD_Japanese-GSDLUW の B-/I-ラベリングでおこない、各トークンを国語研長単位に合わせて組み上げた上で、そこから上三角行列による係り受け解析をおこなった。トークナイザ改造による係り受け解析精度の向上を、表 1「日本語」に示す。評価には UD_Japanese-GSDLUW の test セットを使用し、評価指標は CoNLL 2018 の UPOS / LAS / MLAS を用いた。

Answer.AI が発表した英語 ModernBERT[*4] に対しても、トークナイザの改造を考えた。このトークナイザは「cannot」を 1 語とみなすが、UD_English-EWT は「can」「not」の 2 語として扱う。同様に「don't」を「do」「n't」の 2 語として、「gotta」を「got」「ta」の 2 語として、「wanna」を「wan」「na」の 2 語として扱うことから、これらの単語を狙い撃ちする形で、英語 ModernBERT のトークナイザを改造してみた。トークナイザ改造による係り受け解析精度の向上を、表 1「英語」に示す。評価には UD_English-EWT の test セットを使用した。

ヴュルツブルク大学 Institut für Informatik が発表[*5]したドイツ語 ModernBERT に対しても、トークナイザの改造を考えた。このトークナイザは、縮約冠詞「im」を 1 語とみなすが、UD_German-HDT は「in」「dem」の 2 語として扱う。他の縮約冠詞も同様である。そこで、これらの縮約冠詞を狙い撃ちする形で、ドイツ語 ModernBERT のトークナイザを改造してみた。トークナイザ改造による係り受け解析精度の向上を、表 1「ドイツ語」に示す。評価には UD_German-HDT の test セットを使用した。

Elias Jacob de Menezes Neto[*6]が発表したポルトガル語 ModernBERT に対しても、トークナイザの改造を考えた。このトークナイザは、縮約冠詞「do」を 1 語とみなすが、UD_Portuguese-Bosque は「de」「o」の 2 語として扱う。他の縮約冠詞も同様だが、縮約冠詞の中には「pelo」のように名詞と同型のものがあり、UD_Portuguese-Bosque は縮約冠詞「pelo」を「por」「o」の 2 語として扱うが、名詞「pelo」は 1 語のまま扱う (図 2)。この扱いを実現するため、いったん仮の品詞付与を UD_Portuguese-Bosque のラベリングでおこない、付与された品詞が ADP か DET となった縮約冠詞だけを 2 語に分解してから、あらためて上三角行列による品詞付与と係り受け解析をおこなう、という改造に挑戦した。この改造による係り受け解析精度の向上を、表 1「ポルトガル語」に示す。評価には UD_Portuguese-Bosque の test セットを使用した。

これらのアイデアを援用し、われわれ独自の ModernBERT モデルを、日本語・古典中国語・タイ語について試作した。評価結果を表 2 に示す。なお、本報告書で評価した品詞付与・係り受け解析 ModernBERT モデル (トークナイザ

[*4] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, Iacopo Poli: Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (July 2025), Vol.1: Long Papers, pp.2526-2547.

[*5] https://www.informatik.uni-wuerzburg.de/datascience/news/single/news/moderngbert

[*6] https://docente.ufrn.br/201900343101/perfil

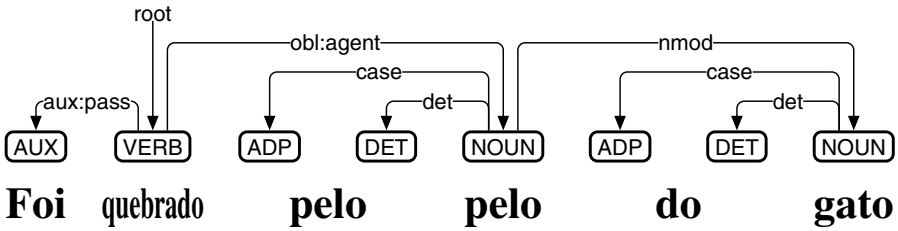


図 2　ポルトガル語における縮約冠詞「pelo」「do」と名詞「pelo」

改造後) は、全て HuggingFace Hub で公開した (表 3)。

6　進捗状況の自己評価と今後の展望

表 1 を見る限り、トークナイザの改造というわれわれの手法は、ModernBERT による品詞付与・係り受け解析においては、かなり有効だといえる。ただし、その有効性を、われわれは独自モデルの製作に活かしきれていない、というのが表 2 の結論である。

学習時の様子を見た限りでは、少なくとも ModernBERT の FillMask モデル作成においては、トークンが長い方が学習効率がいいのだが、それは merges 型トークナイザ (いわゆる GPT 系トークナイザ) に固有の事象にも見える。ならば、FillMask モデルではトークンを長めにとっておいて、品詞付与・係り受け解析モデルでトークナイザを短めにぶった切るのは、どうだろう。このあたりをさらに研究すべく、本共同研究の継続申請をおこなったが、2026 年度は不採択となってしまった。残念だ。

表 1　上三角行列による係り受け解析の評価 (UPOS / LAS / MLAS)

	トークナイザ改造前	トークナイザ改造後
日本語		
sbintuitions/modernbert-ja-30m	48.66 / 22.38 / 12.94	95.88 / 89.20 / 79.89
sbintuitions/modernbert-ja-70m	48.82 / 22.74 / 13,57	96.16 / 89.13 / 80.18
sbintuitions/modernbert-ja-130m	48.79 / 22.58 / 13.48	96.57 / 89.57 / 81.47
sbintuitions/modernbert-ja-310m	48.88 / 22.68 / 13.89	96.62 / 90.38 / 82.62
英語		
answerdotai/ModernBERT-base	95.24 / 88.73 / 80.91	95.97 / 89.52 / 82.04
answerdotai/ModernBERT-large	95.38 / 88.75 / 81.22	96.11 / 89.52 / 82.35
ドイツ語		
LSX-UniWue/ModernGBERT_134M	95.78 / 93.53 / 82.18	98.27 / 96.01 / 84.72
LSX-UniWue/ModernGBERT_1B	94.96 / 91.12 / 78.93	97.44 / 93.63 / 81.42
ポルトガル語		
eliasjacob/ModernBERT-base-portuguese	84.50 / 75.48 / 60.91	95.91 / 85.88 / 70.80
eliasjacob/ModernBERT-large-portuguese	84.49 / 75.39 / 60.51	97.04 / 88.17 / 72.38

表 2　われわれ独自の ModernBERT モデルに対する評価 (UPOS / LAS / MLAS)

日本語	
KoichiYasuoka/modernbert-small-japanese-wikipedia	95.58 / 88.27 / 77.46
KoichiYasuoka/modernbert-base-japanese-wikipedia	95.75 / 88.51 / 78.36
KoichiYasuoka/modernbert-large-japanese-wikipedia	96.42 / 89.56 / 79.94
古典中国語	
KoichiYasuoka/modernbert-small-classical-chinese	91.11 / 77.17 / 73.82
KoichiYasuoka/modernbert-base-classical-chinese	91.24 / 76.33 / 72.99
KoichiYasuoka/modernbert-large-classical-chinese	91.00 / 76.26 / 72.55
タイ語	
KoichiYasuoka/modernbert-base-thai-wikipedia	76.58 / 52.00 / 37.69
KoichiYasuoka/modernbert-large-thai-wikipedia	75.87 / 51.65 / 37.82
KoichiYasuoka/modernbert-base-thai-cc100	80.02 / 62.10 / 48.44

表 3　品詞付与・係り受け解析 ModernBERT モデルの公開

- 日本語 **ModernBERT**
  https://huggingface.co/KoichiYasuoka/modernbert-japanese-30m-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-japanese-70m-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-japanese-130m-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-japanese-310m-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-small-japanese-wikipedia-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-base-japanese-wikipedia-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-large-japanese-wikipedia-ud-embeds

- 古典中国語 **ModernBERT**
  https://huggingface.co/KoichiYasuoka/modernbert-small-classical-chinese-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-base-classical-chinese-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-large-classical-chinese-ud-embeds

- タイ語 **ModernBERT**
  https://huggingface.co/KoichiYasuoka/modernbert-base-thai-wikipedia-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-large-thai-wikipedia-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-base-thai-cc100-ud-embeds

- 英語 **ModernBERT**
  https://huggingface.co/KoichiYasuoka/modernbert-base-english-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-large-english-ud-embeds

- ドイツ語 **ModernBERT**
  https://huggingface.co/KoichiYasuoka/modernbert-german-134m-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-german-1b-ud-embeds

- ポルトガル語 **ModernBERT**
  https://huggingface.co/KoichiYasuoka/modernbert-base-portuguese-ud-embeds
  https://huggingface.co/KoichiYasuoka/modernbert-large-portuguese-ud-embeds