jh241008

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 Unigram 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 GPT2・LLaMA モデルにおいては、トークナイザを単文字化することにより、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。

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

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

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・GPT2・LLaMA 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、RoBERTa を用いた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。

これまでに BERT・RoBERTa・DeBERTa モデルに対して開発してきた手法を援用し、今年度は GPT2・LLaMA モデルへの適用に挑戦したい。GPT2・LLaMA などの言語生成モデルにおいては、各トークンに対する入力ベクトルと出力ベクトルが異なる実装になっている。この点が、これまでわれわれが扱ってきた BERT・RoBERTa・DeBERTa モデルとは大きく違っており、トークナイザ設計においても影響が及ぶ可能性が高い。ただ、GPT2・LLaMA モデルのファインチューニングについては、カスタムインストラクションの手法は多く研究されているものの、係り受け解析については誰もまだ手をつけていない。そもそも、GPT2・LLaMA などの言語生成モデルが、係り受け解析に向いているかどうか (もちろん向いていない可能性もある) すら、誰も調べていない状態である。

われわれとしては、今年度は日本語 (国語研長単位もしくは国語研短単位) を念頭に置きつつ、まずは GPT2・LLaMA による係り受け解析モデルの開発手法そのものを確立したい。

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 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

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ており、RoBERTa による係り受け解析モデルを製作・公開した。

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、RoBERTa・DeBERTa・CamemBERT による係り受け解析用モデルを試作・公開した。

アイヌ語に対しては、カタカナの途中でトークンを切る (たとえば「イタカシ」を「イタク」「アシ」とトークナイズする) 手法を開発し、RoBERTa・DeBERTa による係り受け解析用モデルを試作・公開した。

また、これらの技術をベトナム語に (いわば逆方向に) 援用した。ベトナム語の空白は、単語の区切りではなく音節の区切りであり、したがって、2 音節以上の単語は語中に空白を含む。これら空白を含む単語のうち、古典中国語由来の 2 音節語に対して、空白をまたいでトークンを作成する手法を開発し、BERT・RoBERTa による係り受け解析用モデルを試作・公開した。

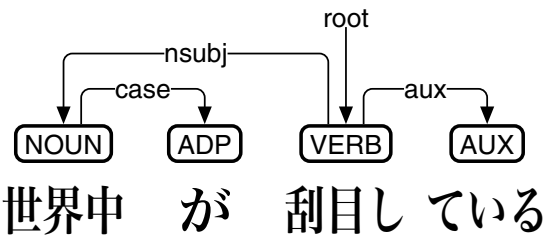
日本語に対しては、字種 (漢字・ひらがな・カタカナ) によってトークン幅を変える必要が

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。

5　今年度の研究成果の詳細

京都大学大学院情報学研究科言語メディア研究室 (ku-nlp) が開発した日本語生成 GPT2 モデルに対し、品詞付与ファインチューニングを試してみたところ、出力側に一段だけ Bellman-Ford アルゴリズムを噛ませる、という工夫を用いることで、かなり高い F1 値を得ることができた。このモデルは単文字トークナイザを採用しており、それが品詞付与において、うまく働いていることがわかった。ただし、GPT2 モデルにおいては、モデル内における情報の流れが文頭から文末への一方向であるために、最終段において逆方向の流れを付与する必要がある、というのが、われわれの知見となった。

この知見を係り受け解析に援用すべく、系列ラベリングによる係り受け解析アルゴリズムを新たに開発した。具体的には、右向きリンクは始点から終点への情報の流れに注目し、左向きリンクは逆向き (終点から始点へ) の情報の流れに注目した上で、係り受け隣接行列を上三角行列に変換する手法である。たとえば、国語研長単位 Universal Dependencies における例文「世界中が刮目している」



であれば、品詞付与・係り受け隣接行列は

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

という正方行列で表せるが、この正方行列を以下の上三角行列に変換する。

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

係り受けラベルにリンクの方向を付与した上で、左向きリンクの各要素を転置する、という手法である。

この手法による品詞付与・係り受け解析を、表 1 に示す日本語 GPT 系モデルに適用した。ファインチューニングには国語研長単位 UD_Japanese-GSDLUW を使用し、評価 (ja_gsdluw-ud-dev.conllu による evaluation)・テスト (ja_gsdluw-ud-test.conllu による predict) をおこなった。評価指標は、CoNLL 2018 の UPOS / LAS / MLAS である。結果を表 2 に示す。

ku-nlp の GPT2 モデルが、パラメータ数に関わらず良い結果を示している。Lightblue 社の karasu-1.1B (LLaMA モデル) が、これに続いている。ku-nlp の GPT2 モデルは単文字トークナイザを用いており、karasu-1.1B は短めのトークンを用いていることから、本手法は、短い文字数のトークンで学習した GPT 系日本語モデルに適している、と考えられる。

表 1　係り受け解析に使用した日本語 GPT 系モデルとその諸元

- https://huggingface.co/ku-nlp/gpt2-small-japanese-char
  GPT2LMHeadModel、86M パラメータ、単文字 GPT2Tokenizer、入出力幅 1024 トークン。
- https://huggingface.co/ku-nlp/gpt2-medium-japanese-char
  GPT2LMHeadModel、296M パラメータ、単文字 GPT2Tokenizer、入出力幅 1024 トークン。
- https://huggingface.co/ku-nlp/gpt2-large-japanese-char
  GPT2LMHeadModel、685M パラメータ、単文字 GPT2Tokenizer、入出力幅 1024 トークン。
- https://huggingface.co/ClassCat/gpt2-base-japanese-v2
  GPT2LMHeadModel、126M パラメータ、非バイト型 GPT2Tokenizer、入出力幅 1024 トークン。
- https://huggingface.co/llm-jp/llm-jp-1.3b-v1.0
  GPT2LMHeadModel、1.23B パラメータ、PreTokenizerFast (llm-jp)、入出力幅 2048 トークン。
- https://huggingface.co/nlp-waseda/gpt2-small-japanese-wikipedia
  GPT2LMHeadModel、106M パラメータ、ReformerTokenizer (Juman++)、入出力幅 1024 トークン。
- https://huggingface.co/nlp-waseda/gpt2-small-japanese
  GPT2LMHeadModel、106M パラメータ、ReformerTokenizer (Juman++)、入出力幅 1024 トークン。
- https://huggingface.co/nlp-waseda/gpt2-xl-japanese
  GPT2LMHeadModel、1.45B パラメータ、ReformerTokenizer (Juman++)、入出力幅 1024 トークン。
- https://huggingface.co/okazaki-lab/japanese-gpt2-medium-unidic
  GPT2LMHeadModel、322M パラメータ、BertJapaneseTokenizer (unidic-lite)、入出力幅 1024 トークン。
- https://huggingface.co/lightblue/karasu-1.1B
  LlamaForCausalLM、992M パラメータ、LlamaTokenizer、入出力幅 2048 トークン。
- https://huggingface.co/yellowback/gpt-neo-japanese-1.3B
  GPTNeoForCausalLM、1.19B パラメータ、GPT2Tokenizer、入出力幅 2048 トークン。
- https://huggingface.co/stockmark/gpt-neox-japanese-1.4b
  GPTNeoXForCausalLM、1.22B パラメータ、GPTNeoXTokenizerFast、入出力幅 1024 トークン。
- https://huggingface.co/cyberagent/open-calm-small
  GPTNeoXForCausalLM、124M パラメータ、GPTNeoXTokenizerFast、入出力幅 2048 トークン。
- https://huggingface.co/cyberagent/open-calm-medium
  GPTNeoXForCausalLM、348M パラメータ、GPTNeoXTokenizerFast、入出力幅 2048 トークン。
- https://huggingface.co/cyberagent/open-calm-large
  GPTNeoXForCausalLM、737M パラメータ、GPTNeoXTokenizerFast、入出力幅 2048 トークン。
- https://huggingface.co/cyberagent/open-calm-1b
  GPTNeoXForCausalLM、1.24B パラメータ、GPTNeoXTokenizerFast、入出力幅 2048 トークン。
- https://huggingface.co/rinna/japanese-gpt-neox-small
  GPTNeoXForCausalLM、118M パラメータ、T5Tokenizer、入出力幅 2048 トークン。
- https://huggingface.co/rinna/japanese-gpt2-xsmall
  GPT2LMHeadModel、36M パラメータ、T5Tokenizer、入出力幅 1024 トークン。
- https://huggingface.co/rinna/japanese-gpt2-small
  GPT2LMHeadModel、106M パラメータ、T5Tokenizer、入出力幅 1024 トークン。
- https://huggingface.co/rinna/japanese-gpt2-medium
  GPT2LMHeadModel、321M パラメータ、T5Tokenizer、入出力幅 1024 トークン。
- https://huggingface.co/rinna/japanese-gpt-1b
  GPT2LMHeadModel、1.21B パラメータ、T5Tokenizer、入出力幅 1024 トークン。

- https://huggingface.co/abeja/gpt2-large-japanese
  GPT2LMHeadModel、717M パラメータ、T5Tokenizer、入出力幅 1024 トークン。
- https://huggingface.co/goldfish-models/jpn_jpan_5mb
  GPT2LMHeadModel、37M パラメータ、AlbertTokenizer、入出力幅 512 トークン。
- https://huggingface.co/goldfish-models/jpn_jpan_10mb
  GPT2LMHeadModel、37M パラメータ、AlbertTokenizer、入出力幅 512 トークン。
- https://huggingface.co/goldfish-models/jpn_jpan_100mb
  GPT2LMHeadModel、119M パラメータ、AlbertTokenizer、入出力幅 512 トークン。
- https://huggingface.co/goldfish-models/jpn_jpan_1000mb
  GPT2LMHeadModel、119M パラメータ、AlbertTokenizer、入出力幅 512 トークン。
- https://huggingface.co/Kendamarron/Tokara-0.5B-v0.1
  Qwen2ForCausalLM、538M パラメータ、Qwen2Tokenizer、入出力幅 32768 トークン。

そこで、表 3 に示す各モデルに対し、単文字トークナイザに置き換える、という改良を試してみることにした。結果として、表 3 の全てにおいて、表 2 に較べ、解析精度の向上が見られた。ただし、それでも解析精度としては今一歩なことから、トークナイズ長を単文字より少し長くする方向が考えられるが、どういうトークナイザが最適なのかは明らかにできなかった。

6　今年度の進捗状況と今後の展望

日本語 GPT 系モデルにおける係り受け解析アルゴリズムを確立した、という点では、十分に評価できるものと信じる。当該アルゴリズムは、ModernBERT など「入力幅の広い」モデルにも援用可能だと考えられることから、今後、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。

表 2　日本語 GPT 系モデルによる国語研長単位係り受け解析の評価 (UPOS / LAS / MLAS)

	評価 (evaluation)	テスト (predict)
ku-nlp/gpt2-small-japanese-char	95.08 / 89.26 / 78.43	**94.78** / **87.85** / 77.25
ku-nlp/gpt2-medium-japanese-char	**95.31** / **90.98** / **81.06**	94.55 / 87.76 / **77.41**
ku-nlp/gpt2-large-japanese-char	94.53 / 88.87 / 78.91	94.06 / 86.91 / 76.64
ClassCat/gpt2-base-japanese-v2	89.97 / 78.07 / 66.45	88.97 / 75.13 / 63.82
llm-jp/llm-jp-1.3b-v1.0	79.29 / 61.52 / 48.95	79.41 / 61.23 / 49.47
nlp-waseda/gpt2-small-japanese-wikipedia	69.58 / 45.86 / 31.35	70.58 / 47.41 / 32.91
nlp-waseda/gpt2-small-japanese	71.05 / 48.52 / 33.43	72.06 / 49.78 / 35.52
nlp-waseda/gpt2-xl-japanese	77.14 / 57.43 / 42.96	77.04 / 56.41 / 41.96
okazaki-lab/japanese-gpt2-medium-unidic	62.03 / 33.97 / 21.20	64.28 / 36.54 / 22.70
lightblue/karasu-1.1B	93.44 / 87.73 / 76.84	93.58 / 86.31 / 75.86
yellowback/gpt-neo-japanese-1.3B	44.99 / 18.44 / 11.00	45.73 / 19.38 / 11.54
stockmark/gpt-neox-japanese-1.4b	43.30 / 17.89 / 10.34	42.65 / 17.11 / 10.13
cyberagent/open-calm-small	41.38 / 16.35 / 8.92	40.83 / 15.45 / 8.52
cyberagent/open-calm-medium	42.50 / 16.50 / 9.02	42.38 / 16.16 / 9.33
cyberagent/open-calm-large	42.43 / 16.72 / 9.33	42.25 / 16.25 / 9.57
cyberagent/open-calm-1b	42.60 / 16.68 / 9.45	42.29 / 16.51 / 9.89
rinna/japanese-gpt-neox-small	55.07 / 29.64 / 18.20	54.06 / 28.25 / 17.79
rinna/japanese-gpt2-xsmall	39.17 / 12.45 / 4.65	40.51 / 12.75 / 4.98
rinna/japanese-gpt2-small	40.60 / 14.03 / 6.00	41.76 / 14.32 / 6.04
rinna/japanese-gpt2-medium	37.94 / 12.40 / 5.28	38.65 / 12.15 / 5.34
rinna/japanese-gpt-1b	40.48 / 16.63 / 9.39	39.89 / 15.05 / 8.61
abeja/gpt2-large-japanese	40.61 / 13.41 / 5.75	42.64 / 14.09 / 6.00
goldfish-models/jpn_jpan_5mb	38.35 / 11.09 / 3.62	39.65 / 12.32 / 4.97
goldfish-models/jpn_jpan_10mb	38.55 / 11.21 / 4.02	39.23 / 11.98 / 4.55
goldfish-models/jpn_jpan_100mb	38.95 / 11.99 / 4.81	40.10 / 12.91 / 5.97
goldfish-models/jpn_jpan_1000mb	39.42 / 12.34 / 5.05	40.52 / 13.26 / 6.46
Kendamarron/Tokara-0.5B-v0.1	69.17 / 46.82 / 32.24	67.32 / 44.26 / 30.79

表 3　トークナイザ改良後の係り受け解析の評価 (UPOS / LAS / MLAS)

	評価 (evaluation)	テスト (predict)
yellowback/gpt-neo-japanese-1.3B	92.79 / 85.08 / 75.17	92.06 / 82.18 / 71.71
stockmark/gpt-neox-japanese-1.4b	94.86 / 88.77 / 80.02	94.22 / 86.50 / 77.98
cyberagent/open-calm-small	94.39 / 86.88 / 75.00	93.54 / 84.36 / 72.72
cyberagent/open-calm-medium	95.22 / 89.12 / 79.36	94.20 / 85.36 / 75.61
cyberagent/open-calm-large	94.86 / 88.84 / 78.82	94.44 / 86.23 / 76.74
cyberagent/open-calm-1b	95.05 / 89.07 / 80.35	94.45 / 86.53 / 78.25
rinna/japanese-gpt-neox-small	89.64 / 77.20 / 64.95	89.39 / 76.72 / 64.94
rinna/japanese-gpt2-xsmall	54.29 / 23.15 / 11.41	56.81 / 25.92 / 13.58
rinna/japanese-gpt2-small	63.20 / 33.45 / 19.73	64.08 / 34.55 / 20.70
rinna/japanese-gpt2-medium	53.06 / 22.75 / 11.68	54.82 / 24.99 / 13.45
rinna/japanese-gpt-1b	95.14 / 89.69 / 81.54	94.62 / 87.52 / 79.38
abeja/gpt2-large-japanese	54.76 / 23.62 / 12.29	57.68 / 26.75 / 14.47
goldfish-models/jpn_jpan_5mb	87.33 / 71.35 / 54.33	86.12 / 68.74 / 52.02
goldfish-models/jpn_jpan_10mb	87.74 / 71.76 / 55.06	86.56 / 70.05 / 53.25
goldfish-models/jpn_jpan_100mb	88.49 / 75.37 / 61.77	88.41 / 74.80 / 60.74
goldfish-models/jpn_jpan_1000mb	89.46 / 77.54 / 65.01	88.93 / 76.52 / 64.08
Kendamarron/Tokara-0.5B-v0.1	94.19 / 88.39 / 77.81	93.58 / 85.79 / 75.41