jh221007

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東アジア人文情報学研究センター）

概要

BERT・RoBERTa・DeBERTa 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 RoBERTa・DeBERTa モデルの開発をおこなっている。古典中国語モデルについては、トークナイザの最大トークン長を大きくしても、係り受け解析の精度がほとんど変化しないことから、単文字トークナイザで十分という結論を得た。日本語モデルについては、カタカナ語に対しては最大トークン長を 4 文字程度にした方が、解析精度が高くなることはわかったが、漢字とひらがなの組み合わせにおいては、最大 4 文字だと解析精度が下がるきらいがあり、字種によるコントロールが必要だと判明した。

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　共同研究分野 (HPCI 資源利用課題のみ)

1.4　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
白須裕之：デジタル処理
鈴木慎吾：コーパス校訂
藤田一乗：コーパス構築

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 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。BERT・RoBERTa・DeBERTa 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、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。

我々が過去に製作した古典中国語係り受け解析システム SuPar-Kanbun は、言語モデルに北京理工大学の GuwenBERT を改造して用いており、漢字 1 文字 1 文字をトークンとみなすような GuwenBERT のトークナイザを流用している。また、我々が過去に製作した日本語係り受け解析システム SuPar-UniDic は、国立国語研究所の UniDic をトークナイザに流用しており、国語研短単位をトークンとみなすような言語モデルを用いている。ただし、これらのトークナイザが文法解析において最適なのかどうか、我々としては確信を得ていない。

我々としては、これまでのようなトークナイザを流用するやり方ではなく、一からトークナイザを設計するやり方で、文法解析に最適な言語モデルを構築したい。しかし、言語モデルの構築には、GPU を長時間稼働して巨大テキストコーパスの学習をおこなう必要があり、しかもトークナイザを変更するごとに一から学習をおこなわなければならない。

本研究課題では、そのようなトークナイザを設計するとともに、それを用いた形態素解析・係り受け解析をおこなうシステムを開発する。

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 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

5 今年度の研究成果の詳細

Transformers の DeBERTa(V2) モデルを、DeBERTa(V2) トークナイザ (Tokenizers の Unigram トークナイザで実装されており、アルゴリズム的には SentencePiece の改良版) と共に、数多く試作した。また、DeBERTa(V2) モデル用に製作した Unigram トークナイザを、Transformers の RoBERTa モデルにも流用することで、DeBERTa(V2)・RoBERTa 間での解析の差異も研究した。

Universal Dependencies 2.10 の古典中国語コーパス UD_Classical_Chinese-Kyoto をもとに、Unigram トークナイザの語彙数 V (`vocab_size`) と最大トークン長 M (`max_piece_length`) をそれぞれ変化させて、古典中国語 DeBERTa(V2) モデルを構築した。CoNLL 2018 の評価指標 (UPOS / LAS / MLAS) で評価・テストした結果を、表 1 に示す。この結果を見る限り、古典中国語モデルについては、トークナイザの最大トークン長を大きくしても、係り受け解析の精度がほとんど変化しない。つまり、語彙数やモデルの規模を考えるならば、古典中国語モデルは、単文字トークナイザで十分ということである。あまり面白くない結果だが、古典中国語 (漢文) という言語が、そもそも漢字 1 文字 1 文字を基本とする言語だということなのかもしれない。

日本語モデルについては、国語研長単位の係り受け解析において、最大トークン長を 4 文字程度にした方が、解析精度が高くなった (表 2)。特に、カタカナ語については明らかに精度

表 1　古典中国語 DeBERTa(V2) トークナイザの比較 (UPOS / LAS / MLAS)

lzh_kyoto-ud-dev.conllu による評価 (evaluation)

	V=8000	**V=16000**	**V=32000**	**V=64000**
M=1	86.96 / 72.86 / 68.01	86.85 / 72.65 / 67.67	86.71 / 72.98 / 67.88	86.88 / 72.63 / 67.86
M=2	86.90 / 72.82 / 67.79	86.94 / 72.84 / 68.02	86.65 / 72.89 / 67.76	86.68 / 72.64 / 67.64
M=4	86.92 / 72.72 / 67.85	86.88 / 72.57 / 67.54	86.54 / 72.75 / 67.85	86.77 / 72.61 / 67.68
M=8	86.73 / 72.42 / 67.56	86.69 / 72.75 / 67.73	86.67 / 72.92 / 67.91	87.04 / 73.04 / 68.12
M=16	86.73 / 72.87 / 67.83	86.95 / 72.81 / 67.93	86.73 / 72.63 / 67.84	86.75 / 72.63 / 67.85

lzh_kyoto-ud-test.conllu によるテスト (predict)

	V=8000	**V=16000**	**V=32000**	**V=64000**
M=1	88.20 / 74.12 / 69.12	88.33 / 74.67 / 69.65	88.07 / 74.41 / 69.13	88.40 / 74.35 / 69.36
M=2	88.40 / 74.71 / 69.58	88.16 / 74.28 / 69.10	88.12 / 74.57 / 69.38	88.43 / 74.81 / 69.69
M=4	88.48 / 74.46 / 69.53	88.34 / 74.77 / 69.45	88.55 / 74.62 / 69.50	88.48 / 74.62 / 69.38
M=8	88.24 / 74.59 / 69.30	88.38 / 74.81 / 69.67	88.38 / 74.71 / 69.41	88.45 / 74.93 / 69.80
M=16	88.37 / 74.56 / 69.21	88.27 / 74.52 / 69.33	88.24 / 74.76 / 69.59	88.39 / 74.79 / 69.69

表 2　日本語 DeBERTa(V2) トークナイザの比較 (UPOS / LAS / MLAS)

ja_gsdluw-ud-dev.conllu による評価 (evaluation)

	V=4000	**V=8000**	**V=16000**	**V=32000**
M=1	71.46 / 57.99 / 35.23	73.40 / 60.59 / 38.56	73.51 / 61.78 / 39.01	71.29 / 56.86 / 34.30
M=2	76.64 / 63.20 / 41.88	77.73 / 64.10 / 43.21	77.80 / 64.83 / 43.73	78.66 / 66.17 / 45.00
M=4	75.37 / 60.53 / 38.57	80.30 / 69.16 / 48.07	80.52 / 69.43 / 48.82	78.60 / 65.19 / 44.62
M=8	79.72 / 68.96 / 46.51	80.67 / 69.27 / 48.86	80.45 / 68.85 / 47.70	79.74 / 67.48 / 46.59
M=16	78.51 / 66.62 / 44.59	78.75 / 65.44 / 44.67	80.86 / 69.59 / 49.23	80.94 / 70.72 / 49.53

ja_gsdluw-ud-test.conllu によるテスト (predict)

	V=4000	**V=8000**	**V=16000**	**V=32000**
M=1	68.93 / 54.91 / 32.20	71.22 / 58.26 / 35.29	70.72 / 58.59 / 35.75	68.67 / 53.94 / 31.92
M=2	74.03 / 59.49 / 37.59	75.34 / 60.50 / 39.13	75.22 / 60.57 / 38.97	76.07 / 62.40 / 40.99
M=4	73.29 / 57.97 / 35.67	78.82 / 66.46 / 45.12	78.50 / 66.55 / 44.87	76.06 / 61.89 / 40.14
M=8	77.43 / 67.14 / 44.34	78.69 / 66.20 / 44.86	78.17 / 65.34 / 44.50	77.93 / 64.80 / 43.68
M=16	76.51 / 65.19 / 42.64	76.74 / 62.58 / 41.40	79.10 / 67.01 / 45.92	79.03 / 66.94 / 45.73

が良くなった。ただ、漢字とひらがなの組み合わせにおいては、最大 4 文字だと解析精度が下がるきらいがあり、どうやら字種によるコントロールが必要だと考えられる。

一方、韓国語モデルとアイヌ語モデルについては、文字より小さな単位でトークナイザを設計した方が、解析精度が上がる場合がある、ということが判明した。韓国語については、用言の活用が文字の中で起こる (末子音が変化したり、母音の挿入が行われる) ため、用言に対してはハングルではなく、字母のレベルでトークナイザを設計する方が解析精度が上がる。一方、アイヌ語については、末子音と母音のアンシェヌマンにより、カタカナの途中に単語の切れ目が来ることがあり、そのような場合はカタカナの途中でトークンを切るしかない。

すなわち、言語によっては文字より小さな単位を考える必要がある、という点までは明らかとなった。しかしながら、どのような場合において文字単位 (あるいは複数の文字単位) が良く、どのような場合において文字より小さな単位が良いのかは、まだ判然としていない。

6　今年度の進捗状況と今後の展望

古典中国語モデルについては、まずは予定通りの進捗状況と言える。日本語モデルについても、予定通りの進捗状況ではあるのだが、韓国語・アイヌ語に関する (いわば予定外の) 知見が得られたことから、文字より小さい単位による検討を必要としている。その意味で日本語モデルは、予定と異なる方向に進んでいる気はするものの、用言の活用という特殊な体系を有する言語として、様々な手法によるトークナイザを、今後どんどん試していく必要があると考えられる。

7　研究業績一覧（発表予定も含む）

学術論文 （査読あり）

- Yasuoka Koichi: Universal Dependencies와 BERT/RoBERTa 모델을 통한 고전 중국어 정보처리, Journal of Applied Studies on Sinograph and Literary Sinitic, Vol.1 (December 2022), pp.127-163.

国際会議プロシーディングス　（査読あり）

国際会議発表　（査読なし）

- Koichi Yasuoka: Reed-Kellogg, Tesnière, Мельчук, and Universal Dependencies, Towards a comprehensive collaborative research environment for the study of premodern Chinese culture (May 27, 2022).
- 安岡孝一, 安岡素子: 古典中国語の形態素解析と係り受け解析, 근역한문학회 2022년 추계 기획학술대회: 디지털과 한문 고전 연구 (2022 年 11 月), pp.148-160.

国内会議発表　（査読なし）

- 安岡孝一: 青空文庫 DeBERTa モデルによる国語研長単位係り受け解析, 東洋学へのコンピュータ利用, 第 35 回研究セミナー (2022 年 7 月 29 日), pp.29-43.
- 安岡孝一: BERT/RoBERTa/DeBERTa モデルによる形態素解析と係り受け解析, データ活用社会創成シンポジウム 2022 (2022 年 12 月 20 日).
- 安岡孝一: ローマ字・カタカナ・キリル文字併用アイヌ語 RoBERTa・DeBERTa モデルの開発, 情報処理学会研究報告, Vol.2023-CH-131『人文科学とコンピュータ』, No.7 (2023 年 2 月 18 日), pp.1-7.

公開したライブラリ等

- Koichi Yasuoka:
  https://pypi.org/project/esupar

その他（特許，プレス発表，著書等）

- 安岡孝一: Universal Dependencies と BERT/RoBERTa/DeBERTa モデルによる多言語情報処理 (2022 年 12 月版), 京都大学人文科学研究所・未踏科学研究ユニット・データサイエンスで切り拓く総合地域研究ユニット.
  http://hdl.handle.net/2433/278350