jh231002

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 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。タイ語モデルについては、タイ文字クラスターと矛盾しないよう工夫しつつ、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がわかった。アイヌ語モデルについては、カタカナ・キリル文字をローマ字に変換した上で、最大トークン長を 8 文字程度にした方が、解析精度が高くなることがわかった。

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　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
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 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。

我々の研究グループは、多言語係り受け解析システム esupar [6] を開発中である。esupar は、古典中国語 (漢文) の解析に際し、漢字 1 文字 1 文字をトークンとみなすような言語モデルを用いている。この単文字トークナイザが、古典中国語の文法解析において最適であることは、2022 年度の jh221007 において明らかにできた。一方、日本語の解析に際しては、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。しかし、このトークナイザが、近代日本語の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、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。今年度は、日本語 (国語研長単位) の係り受け解析モデルにおけるトークナイザを、何とか最適化したい。また、アイヌ語・韓国語・ベトナム語のような、空白と単語の区切りが一致しない書写言語についても、良いトークナイザと係り受け解析エンジンの開発に着手したい。

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いる。

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

日本語向けトークナイザの設計は、正直なところ、なかなかうまく進んでいない。漢字・ひらがな・カタカナが混在する日本語においては、字種それぞれにパラメータが違うだろうと予想していたものの、特にひらがなの動きが複雑で、漢字の直後にある場合でも、それが名詞につく助詞なのか、動詞や形容詞に含まれる活用語尾なのかによって、「切れ目」の位置が変わってくる。しかも現代日本語と近代日本語で、活用語尾の形式が異なる。現代仮名遣いと歴史的仮名遣いの差のみならず、活用語尾そのものが違う上に、音便も異なっている。さらに上代日本語では、ひらがなが発明されておらず、全てが漢字で書かれている [1] 上に、音節の種類数

が近代日本語より多い。上代日本語での活用形式が、近代日本語へと圧縮された上で、現代日本語へと引き継がれるはずなのだが、トークナイザの設計という局面においては、そう簡単な話では無いということである。

タイ語トークナイザについては、Unigram トークナイザの最大文字長を 4 文字 (母音や声調記号を 1 文字に数える) に限定する、というあたりが、孤立語であるタイ語の係り受け解析に有効である [2]。ただし、最大 4 文字なら何でもいいというわけではなく、タイ文字クラスター (**คลัสเตอร์อักษรไทย**) と矛盾しないよう、トークンの取捨選択をおこなう ([8] 図 70) 必要があった。このような形で作成したタイ語トークナイザを用いて、タイ語 DeBERTa モデルを製作し、esupar のタイ語係り受け解析モジュールとして公開した。

アイヌ語トークナイザについては、カタカナの途中に語境界が来る場合を、どう処理するかが問題となった。たとえば「オカヤン」は「オカイ」「アン」がアンシェヌマンを起こしている例 [4] だし、「コトマン」は「コトム」「アン」である [7]。この問題を解決すべく、カタカナ・キリル文字をローマ字に変換するモジュールを esupar に実装した。さらに変換後のローマ字において、Unigram トークナイザの最大文字長を 8 文字に限定する、という手法 ([8] 図 171) で、アイヌ語トークナイザを設計した。このような形で作成したアイヌ語トークナイザを用いて、アイヌ語 RoBERTa モデルを製作し、esupar のアイヌ語係り受け解析モジュールとして公開した。

また、これらのトークナイザ技術を、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 [2] し、ベトナム語係り受け解析モデル `https://huggingface.co/KoichiYasuoka/phobert-base-vietnamese-ud-goeswith` として公開した。ただし、3 音節以上の単語をどう扱うかについては、まだ手探りの状態である。

6　今年度の進捗状況と今後の展望

タイ語トークナイザについては、ほぼ期待した成果が得られたと言える。アイヌ語トークナイザについては、もう少しチューニングの必要があるとは思われるものの、だいたいの方向性は見えている。日本語トークナイザに関しては、まだまだ期待した成果は得られておらず、今後も研究を継続したい。できれば、現代中国語や韓国語にも着手したいが、道程は長そうである。

GPT・LLaMA などの言語生成モデルにおけるトークナイザについても、我々としては気になるところである。ただ、言語生成モデルにおけるトークナイザの性能を、どういう指標で測るべきなのかについては、現時点の我々も五里霧中である。

7　研究業績一覧（発表予定も含む）

学術論文 （査読あり）

[1] 安岡孝一, ウィッテルン クリスティアン, 池田巧, 藤田一乗, 山崎直樹, 二階堂善弘, 鈴木慎吾, 守岡知彦, 師茂樹: 『日本書紀』 Universal Dependencies への挑戦, 人文科学とコンピュータシンポジウム「じんもんこん 2023」論文集 (2023 年 12 月), pp.169-176.

国際会議プロシーディングス　（査読あり）

[2] Koichi Yasuoka: Sequence-Labeling RoBERTa Model for Dependency-Parsing in Classical Chinese and Its Application to Vietnamese and Thai, ICBIR 2023: 8th International Conference on Business and Industrial Research (May 2023), pp.169-173.

国際会議発表　（査読なし）

国内会議発表　（査読なし）

[3] 安岡孝一: BERT/RoBERTa/DeBERTa モデルによる多言語係り受け解析, SS 研 HPC フォーラム 2023「自然言語処理と高性能計算〜シナジーを探る〜」(2023 年 8 月 21 日).

[4] 安岡孝一, 安岡素子: ローマ字・カタカナ・キリル文字によるアイヌ語 Universal Dependencies の可能性, Evidence-based Linguistics Workshop 2023 (2023 年 9 月 15 日), pp.47-60.

[5] 安岡孝一: 屈折語・孤立語・膠着語・抱合語そして Universal Dependencies, 東西学術研究所 / 経済・政治研究所 / 法学研究所 3 研究所合同シンポジウム「AI 社会の現在」 (2023 年 11 月 25 日).

公開したライブラリ等

[6] https://pypi.org/project/esupar

その他（特許，プレス発表，著書等）

[7] 安岡孝一, 安岡素子: Universal Dependencies で読むアイヌ語訳『五倫名義解』, 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研究年報 2023 (2023 年 9 月).
http://hdl.handle.net/2433/286023

[8] 安岡孝一: Universal Dependencies と BERT/RoBERTa/DeBERTa モデルによる多言語情報処理 (2023 年 11 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/2433/286274