様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ４ 日現在

機関番号：１４３０１
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０４４８１

研究課題名（和文）古典漢文依存文法コーパスにもとづく係り受け構造の自動抽出

研究課題名（英文）Dependency-Parsing in Classical Chinese under Universal Dependencies

研究代表者
安岡　孝一（YASUOKA, Koichi）

京都大学・人文科学研究所・教授

研究者番号：２０２３０２１１

交付決定額（研究期間全体）：（直接経費）　　13,400,000円

研究成果の概要（和文）：古典漢文の白文（単なる漢字の列）に対し、文切り、単語の組み上げ、品詞付与、単語間の係り受け解析、節の組み上げ、節間の係り受け解析、をおこなう手法を開発した。この手法を、言語モデルRoBERTa-Classical-Chineseをチューニングする形で、実装・公開をおこなった。また、この手法が、ベトナム語やタイ語にも適用可能であったことから、同様に実装・公開をおこなった。

研究成果の学術的意義や社会的意義
学術的意義としては、古典漢文の白文（単なる漢字の列）が、本研究の手法により、文・節・単語の単位に区切ることが出来るようになる上に、それらの関係（どの単語が動詞で、その主語や目的語はどれなのか、など）が、非常に高い精度で自動解析できるようになった。
一方、社会的意義としては、本研究の手法が、ベトナム語やタイ語にも適用可能であるという点が挙げられる。ベトナム語もタイ語も、単語の切れ目すら見極めるのが難しい言語であり、それが自動解析できるようになる意義は大きい。

研究成果の概要（英文）：We have developed RoBERTa-Classical-Chinese and its fine-tuned models for Classical Chinese to perform sentence segmentation, word tokenization, part-of-speech tagging, dependency-parsing between words, phrase detection, and dependency-parsing between phrases. And we have applied our methods to other isolating languages, such as Vietnamese and Thai.

研究分野：人文情報学

キーワード：言語処理　古典漢文

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

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は、その前身である附属東洋学文献センター時代から、現代に至るまで、約 130,000 タイトルの古典漢籍文献を収集し、その保存と公開につとめてきた。また、1980 年代から、京都大学大型計算機センター（現、京都大学学術情報メディアセンター）との共同研究で、古典漢籍の全文テキストデータベース化をおこなってきた。

これらの膨大な古典漢文テキストをコンピュータで処理するためには、白文（単なる漢字の列）ではなく、テキストを自然言語解析する必要がある。古典漢文のように、単語の間にも文の間にも区切りを持たない書写言語の解析では、まず、単語を認識することが必須であり、形態素解析を十全におこなった上で、その結果を元に構文解析を進めていく、という手法を取らざるを得ない。ただし、現代中国語と違って、単語の間にも文の間にも区切りを持たない古典漢文に対しては、現代中国語の解析手法が無力であり、新たな手法を開発しなければならないという問題があった。

この問題に対し、研究代表者は、2008 年度より京都大学人文科学研究所共同研究班「東アジア古典文献コーパスの研究」を組織し、古典漢文に対する形態素解析の研究を開始した。この共同研究班において、われわれは、言語に依存しない解析エンジンとして MeCab を選び、さらに古典漢文を形態素解析するための品詞分類を研究した。また、この共同研究班および後身の共同研究班「東アジア古典文献コーパスの応用研究」を母体として、2010～2012 年度に科学研究費基盤研究(B)『形態素解析のための品詞情報つき古典漢文コーパスの構築』、2013～2015 年度に科学研究費基盤研究(B)『品詞素性情報つき古典漢文コーパスの発展的応用』により、古典漢文コーパスの構築と形態素解析の研究をおこなった。

これに続く共同研究班「東アジア古典文献コーパスの実証研究」において、われわれは、古典漢文に対する構文解析手法の研究を開始した。さらに、この共同研究班を母体として、2017～2019 年度に科学研究費基盤研究(B)『古典漢文形態素コーパスにもとづく動詞の作用域の自動抽出』により、古典漢文における動賓構造の自動抽出に精力を傾注した。多数の構文解析手法を比較・検討した結果、われわれは、Мельчук の依存文法(Dependency Grammar)が古典漢文における動賓構造の自動抽出に適している、という結論を得た。また、コンピュータによる解析という視点から見た場合、古典漢文の動賓構造に対しては、いわゆる SVO (Subject-Verb-Object)構造ではなく、句末に終助詞を伴う動賓終(predicate-object-final)構造を、基本構造として扱うのが適切である、という結論を得た。

これらの結論にもとづき、われわれは、Мельчук の依存文法をデジタル向けに改良した Universal Dependencies という記述手法により、『四書』（孟子、論語、大學、中庸）の依存文法コーパスを Treebank として制作した。さらに、チェコ語の依存文法解析エンジンである UDPipe を改造して、われわれの古典漢文 Treebank と組み合わせたところ、古典漢文における動詞の作用域は、かなり高い精度で自動抽出できるようになった。

２．研究の目的

では、われわれのこの手法により、古典漢文の文法解析は、どの程度まで可能となるのか。Мельчук の依存文法は、動賓構造の自動抽出や、単語と単語の間の係り受け解析には有効だが、それより高次（節レベルあるいは文レベル）の関係解析に拡張可能なのか。拡張可能であるならば、どのような手法を開発すれば自動解析できるようになるのか。これが本研究の目的である。

３．研究の方法

古典漢文に対し形態素解析と依存文法解析をおこなった上で、単語と単語の間の係り受け関係、節と節の間の係り受け関係、文と文の間の係り受け関係を、自動抽出する手法の構築をおこなう。この手法を構築するために、各レベル（単語・節・文）での係り受け関係を記述するための文法と、その文法にもとづく Treebank の構築を並行しておこない、単語・節・文の順に、係り受け関係を自動抽出する手法を完成する。

４．研究成果

MeCab と UDPipe を組み合わせたわれわれの従来手法では、本研究の目標に遠く及ばず、解析アルゴリズム全体の再設計をおこなった。アイデアとしては google が英語向けに開発した BERT (Bidirectional Encoder Representations from Transformers)を古典漢文に適用することを考えたものの、そのままではうまくいかなかった。BERT は、単語や文の間に区切りがある言語を想定して設計されており、単語の間にも文の間にも区切りを持たない古典漢文には適用が難しい。様々な言語モデルを試した結果、Facebook AI が開発中だった RoBERTa から「単語」の概念を捨て去り、代わりに漢字 1 文字 1 文字をトークンとみなしたモデルを設計して、古典漢文向けの言語モデル RoBERTa-Classical-Chinese (base モデルおよび large モデル) を製作した。

この RoBERTa-Classical-Chinese を系列ラベリングでファインチューニングする形で、品詞付与と単語組み上げを同時におこなう手法を開発した。ファインチューニングには、われわれが製作した『四書』依存文法コーパス Treebank を拡張して用いた。また、『四書』依存文法コーパス Treebank の拡張に際し、白文を節ないしは文の単位で切る手法を系列ラベリング上に展開するアルゴリズムを開発した。単語間の係り受け解析については、スタンフォード大学で Universal Dependencies 向けに開発された Biaffine アルゴリズムを借りることにした。これらの手法をまとめて、言語解析モジュール SuPar-Kanbun として発表した。

古典漢文における単語間の係り受け解析アルゴリズムを、節の係り受けへと拡張するにあたって、RoBERTa-Classical-Chinese のファインチューニング手法を見直すことにした。白文の各漢字をグラフアルゴリズムにおけるノードとみなすならば、節への組み上げは、Bellman-Ford アルゴリズムを系列ラベリングに適用する形で、節を表すメタノードへの組み上げとして実装できる。ならば、メタノード間の係り受け解析は、Chu-Liu-Edmonds アルゴリズムにおいて「ノードをまとめる」ロジックを、メタノードへの組み上げにも援用すれば、ノード間リンクの「重み」がメタノード間リンクへと援用できるはずである。

これらの手法をまとめあげ、RoBERTa-Classical-Chinese の系列ラベリングを用いて、単語の組み上げ、品詞付与、単語間の係り受け解析、節の組み上げ、節間の係り受け解析、を同時におこなえる手法を開発した。具体的には、ノード（漢字）間リンクの「重み」を、係り受けグラフにおける隣接確率の対数オッズ比にすることで、Bellman-Ford アルゴリズムにも Chu-Liu-Edmonds アルゴリズムにも、「重み」付き隣接行列として適用できるよう工夫した。また、この手法が、古典漢文のみならず、ベトナム語やタイ語にも適用可能であることを、様々な手法の比較実験中に（たまたま）発見した。ベトナム語やタイ語は、古典漢文の影響を強く受けており、しかも孤立語である。ただし、同様に古典漢文の影響を受けているはずの現代中国語や近代日本語には、この手法はあまりうまく適用できなかった。

ここまでで製作した単語間・節間の係り受け解析アルゴリズムを、文の係り受けへと拡張するにあたって、われわれは、先に開発した「白文を文の単位で切る手法」を融合し、メタノードを文に拡張する手法を試してみた。しかし、この手法は残念ながらうまくいかなかった。「重み」付き隣接行列の内部を手作業で検討してみたところ、元々のわれわれの仮定が甘かったことを、われわれは思い知ることになった。すなわち、漢字の集まりが単語であり、単語の集まりが節である、というところまでは正しいものの、節の集まりが文である、というのは必ずしも正しくないらしい。もちろん表層的には、節の集まりが文を構成している。しかし、文と文の関係を考えた場合には、そのような表層的なモデリングではなく、むしろ、白文全体（いわば段落）を区切る単位が文なのであり、白文全体と文の関係を射影する形で、文と文の関係が現れてくるようなのである。

そうすると、RoBERTa-Classical-Chinese ではなく、古典漢文向けの新たな言語モデルを設計しなければ、文と文の係り受け解析を高い精度でおこなうのは難しそうだ。しかし、Sentence BERT など、いくつか既存の言語モデルを古典漢文に適用してみたものの、RoBERTa-Classical-Chinese での係り受け解析精度にすら及ばなかった。これを言い換えると、Мельчук の依存文法は、節と節の係り受けへは拡張可能だが、文と文の係り受けには必ずしも拡張できない、というのが、古典漢文に関する現時点でのわれわれの結論である。

５．主な発表論文等

〔雑誌論文〕　計11件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　11件）

１．著者名 Yasuoka Koichi	４．巻 ICBIR 2023
２．論文標題 Sequence-Labeling RoBERTa Model for Dependency-Parsing in Classical Chinese and Its Application to Vietnamese and Thai	５．発行年 2023年
３．雑誌名 8th International Conference on Business and Industrial Research	６．最初と最後の頁 169-173
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ICBIR57571.2023.10147628	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一，ウィッテルン クリスティアン，守岡知彦，池田巧，山崎直樹，二階堂善弘，鈴木慎吾，師茂樹，藤田一乘	４．巻 63
２．論文標題 古典中国語（漢文）Universal Dependenciesとその応用	５．発行年 2022年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 355-363
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一	４．巻 1
２．論文標題 Universal DependenciesとBERT/RoBERTaモデルによる古典中国語情報処理 (in Korean)	５．発行年 2022年
３．雑誌名 Journal of Applied Studies on Sinograph and Literary Sinitic	６．最初と最後の頁 127-163
掲載論文のDOI（デジタルオブジェクト識別子） 10.26523/HERC.2022.1.127	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一，安岡素子	４．巻 2022
２．論文標題 古典中国語の形態素解析と係り受け解析	５．発行年 2022年
３．雑誌名 槿域漢文学会2022年秋季企画学術大会	６．最初と最後の頁 171-183
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 二階堂善弘	4．巻 2022
2．論文標題 画像とテキストの位置づけ	5．発行年 2022年
3．雑誌名 KU-ORCASが開くデジタル化時代の東アジア文化研究	6．最初と最後の頁 123-130
掲載論文のDOI（デジタルオブジェクト識別子） 10.32286/00026586	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 安岡孝一	4．巻 2021
2．論文標題 Transformersを用いた古典中国語(漢文)文切りモデルの製作	5．発行年 2021年
3．雑誌名 人文科学とコンピュータシンポジウム「じんもんこん2021」論文集	6．最初と最後の頁 104-109
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 守岡知彦	4．巻 33
2．論文標題 CHISEのWeb API化の試み、ついでに、RDF化四度目の正直？	5．発行年 2021年
3．雑誌名 東洋学へのコンピュータ利用	6．最初と最後の頁 69-87
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 安岡孝一	4．巻 2020
2．論文標題 Universal Dependenciesにもとづく多言語係り受け可視化ツールdeplacy	5．発行年 2020年
3．雑誌名 人文科学とコンピュータシンポジウム「じんもんこん2020」論文集	6．最初と最後の頁 95-100
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安岡孝一	４．巻 33
２．論文標題 TransformersのBERTは共通テスト『国語』を係り受け解析する夢を見るか	５．発行年 2021年
３．雑誌名 東洋学へのコンピュータ利用	６．最初と最後の頁 3-34
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Tomohiko Morioka	４．巻 Part II
２．論文標題 Viewpoints on the Structural Description of Chinese Characters	５．発行年 2020年
３．雑誌名 Grapholinguistics in the 21st Century―2020	６．最初と最後の頁 683-712
掲載論文のDOI（デジタルオブジェクト識別子） 10.36824/2020-graf-mori	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Christian Wittern	４．巻 33
２．論文標題 Kanripo X: A tagset for connecting digital texts	５．発行年 2021年
３．雑誌名 東洋学へのコンピュータ利用	６．最初と最後の頁 35-67
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計3件（うち招待講演　3件／うち国際学会　1件）

１．発表者名 安岡孝一
２．発表標題 古典中国語の形態素解析と係り受け解析
３．学会等名 橿域漢文学会2022年秋季企画学術大会（招待講演）（国際学会）
４．発表年 2022年

１．発表者名
安岡孝一

２．発表標題
漢字・漢語・漢文の言語情報処理

３．学会等名
日本ソフトウェア科学会第38回大会（招待講演）

４．発表年
2021年

１．発表者名
安岡孝一

２．発表標題
世界のUniversal Dependenciesと係り受け解析ツール群

３．学会等名
第3回Universal Dependencies公開研究会（招待講演）

４．発表年
2021年

〔図書〕　計1件

１．著者名 安岡孝一	４．発行年 2022年
２．出版社 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	５．総ページ数 101
３．書名 Universal DependenciesとBERT/RoBERTa/DeBERTaモデルによる多言語情報処理(2022年12月版)	

〔産業財産権〕

〔その他〕

「古典中国語のコーパスの研究」共同研究班ログ
http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/archive2023.html

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	山崎　直樹 (Yamazaki Naoki) (30230402)	関西大学・外国語学部・教授 (34416)	
研究分担者	二階堂　善弘 (Nikaido Yoshihiro) (70292258)	関西大学・文学部・教授 (34416)	
研究分担者	師　茂樹 (Moro Shigeki) (70351294)	花園大学・文学部・教授 (34313)	
研究分担者	Ｗｉｔｔｅｒｎ　Ｃ. (Wittern Christian) (20333560)	京都大学・人文科学研究所・教授 (14301)	
研究分担者	池田　巧 (Ikeda Takumi) (90259250)	京都大学・人文科学研究所・教授 (14301)	
研究分担者	守岡　知彦 (Morioka Tomohiko) (40324701)	京都大学・人文科学研究所・助教 (14301)	
研究分担者	白須　裕之 (Shirasu Hiroyuki) (30828570)	京都大学・人文科学研究所・助教 (14301)	
研究分担者	鈴木　慎吾 (Suzuki Shingo) (20513360)	大阪大学・言語文化研究科（言語社会専攻、日本語・日本文化専攻）・准教授 (14401)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
チェコ	カレル大学			
米国	スタンフォード大学			
中国	北京理工大学	南京農業大学		
その他の国・地域（台湾）	東呉大学			