古典中国語 (漢文) ModernBERT の開発

安岡孝一*

1　はじめに

2024 年 12 月 19 日に Answer.AI が発表した ModernBERT [a] は、入出力幅 (max position embeddings) 8192 トークンの言語モデルを 1.49 億 (149M) パラメータで実現する、という途方もないものだった。筆者が研究代表者を務める学際大規模情報基盤共同利用・共同研究拠点公募型共同研究『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』(共同研究者: 山崎直樹・二階堂善弘・師茂樹・鈴木慎吾・Christian Wittern・池田巧・守岡知彦・李媛・劉冠偉) では、これまでに多種多様な係り受け解析エンジンを開発してきた [1, 2, 3, 4] が、入出力幅は 512 トークンが中心であり、その 16 倍もの入出力幅は経験がなかった。入出力幅が拡がれば、新たな解析アルゴリズムの可能性が生まれる。

係り受け解析での隣接確率行列を考えると、8192 トークンもあれば 90×90 の正方行列が、そのままモデルに乗ってしまう。三角行列に圧縮できれば、126×126 までは乗りそうである。つまり、隣接確率行列をモデルに乗せてしまった形での解析アルゴリズムを、開発可能だということである。

本稿では、古典中国語 (漢文) ModernBERT モデルの開発をおこないつつ、古典中国語 Universal Dependencies [1] を題材に、ModernBERT での係り受け解析アルゴリズムの可能性を探る。

2　Universal Dependencies の概要

Universal Dependencies (UD) は、書写言語における品詞・形態素属性・依存構造 (係り受け関係) を、言語に関わらず記述する手法 [5] である。句構造を考慮せずに係り受け関係を記述することで、言語横断性を高めており、全ての文法構造を単語間のリンクで記述するのが特徴である。

依存構造解析それ自体は、Tesnière の構造的統語論 [6] に源を発し、Мельчук の有向グラフ記述 [7] によって、一応の完成を見た手法である。その最大の特長は、いわゆる動詞中心主義によって言語横断的な記述が可能だという点にあり、Мельчук 依存文法をコンピュータ向けに洗練した UD においても、言語に関わらない記述、という特長が前面に押し出されている。UD における文法構造記述は、句構造を考慮せず、全てを単語間のリンクとして表現する。これにより、言語横断的な文法構造記述を可能としている。

UD 係り受けコーパスの交換用フォーマットとして、CoNLL-U と呼ばれるタブ区切りテキスト (文字コードは UTF-8) が規定されている。CoNLL-U の各行は各単語に対応しており、表 1 に示す 10 個のタブ区切りフィールドで構成される。ID・FORM・LEMMA は、単語そのものに関するフィールドである。UPOS・XPOS・FEATS は、単語の品詞と

*京都大学人文科学研究所附属人文情報学創新センター

[a] https://huggingface.co/blog/modernbert

表 1: CoNLL-U の各フィールド

1. ID: 単語ごとに付与されたインデックスで、文ごとに 1 から始まる整数。縮約語に対しては、単語の範囲を示すのも可。
2. FORM: 語、または、句読記号。
3. LEMMA: 基底形、語幹。
4. UPOS: UD で規定された言語普遍な品詞タグ (表 2)。
5. XPOS: 言語固有の品詞タグ。
6. FEATS: UD で規定された言語普遍な形態素属性のリスト。言語固有の拡張も可。
7. HEAD: 当該の単語の係り受け元 ID。係り受け元が無い場合は 0 とする。
8. DEPREL: UD で規定された言語普遍な係り受けタグ (表 3)。HEAD が 0 の場合は root とする。言語固有の拡張も可。
9. DEPS: 複数の係り受け元を持つ場合、全ての HEAD:DEPREL ペア。
10. MISC: その他のアノテーション。

表 2: UD 品詞タグ (UPOS)

Open class words	**Closed class words**	**Other**
ADJ 形容詞	ADP 側置詞	PUNCT 句読点
ADV 副詞	AUX 助動詞	SYM 記号
INTJ 感嘆詞	CCONJ 並列接続詞	X その他
NOUN 名詞	DET 限定詞	
PROPN 固有名詞	NUM 数詞	
VERB 動詞	PART 接辞	
	PRON 代名詞	
	SCONJ 従属接続詞	

表 3: UD 係り受けタグ (DEPREL)

	Nominals	**Clauses**	**Modifier Words**	**Function Words**
Core arguments	nsubj 主語 obj 目的語 iobj 間接目的語	csubj 節主語 ccomp 節目的語 xcomp 節補語		
Non-core dependents	obl 斜格補語 vocative 呼称語 expl 形式語 dislocated 外置語	advcl 連用修飾節	advmod 連用修飾語 discourse 談話要素	aux 動詞補助成分 cop 繋辞 mark 標識
Nominal dependents	nmod 体言による連体修飾語 appos 同格 nummod 数量による修飾語	acl 連体修飾節	amod 用言による連体修飾語	det 決定語 clf 類別語 case 格表示

Coordination	**MWE**	**Loose**	**Special**	**Other**
conj 接続 cc 接続語	fixed 固着 flat 並列 compound 複合	list 細目 parataxis 隣接表現	orphan 親なし goeswith 泣き別れ reparandum 言い損じ	punct 句読点 root 親 dep 未定義

形態素属性に関するフィールドである。HEAD・DEPREL・DEPS は、単語の係り受けに関するフィールドである。

UD における係り受け関係は、単語間の有向グラフを HEAD と DEPREL で記述する。HEAD は、その単語に入る有向枝のリンク元 ID を示しており、DEPREL は、その有向枝における係り受けタグである。ただし、HEAD が 0 の場合、その枝に入るリンク元は存在しない。リンクの本数は単語の個数に等しく、各リンクのリンク先は、全て互いに異なっている。すなわち、各単語から出るリンクは複数の可能性があるが、各単語に入るリンクは 1 つだけである。なお、リンクはループしない。

UD の係り受けリンクは、Мельчук 依存文法の後裔にあたり、いわゆる動詞中心主義である。動詞をリンク元として、主語や目的語へとリンクする。修飾関係においては、被修飾語から修飾語へとリンクする。ただし、側置詞 (前置詞や後置詞) を体言の修飾語だとみなす点 [8] が、Мельчук とは異なっている。また、コピュラ文においては動詞中心主義を採らず、補語をリンク元として、主語や繋辞へとリンクする。

古典中国語 UD の例として、「孟子見梁惠王」の CoNLL-U と、deplacy [9] による可視化を図 1 に示す。ただし、本稿のアルゴリズムでは、XPOS・DEPS は使用していない。

# ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	孟子	孟子	PROPN	n, 名詞, 人, 複合的人名	_	2	nsubj	_	SpaceAfter=No
2	見	見	VERB	v, 動詞, 行為, 動作	_	0	root	_	SpaceAfter=No
3	梁	梁	PROPN	n, 名詞, 主体, 国名	_	5	nmod	_	SpaceAfter=No
4	惠	惠	PROPN	n, 名詞, 人, その他の人名	_	5	compound	_	SpaceAfter=No
5	王	王	NOUN	n, 名詞, 人, 役割	_	2	obj	_	SpaceAfter=No

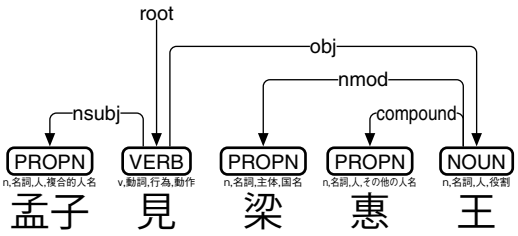


図 1: 「孟子見梁惠王」の CoNLL-U と可視化

3 古典中国語 ModernBERT の製作

漢籍リポジトリ (Kanripo)[10] の繁體字テキスト 6 億字を、単文字トークナイズした上で、Transformers [11] version 4.51.3 を用いて、古典中国語 ModernBERT 繁體字版 small・base・large モデルを製作した。製作には、mdx の NVIDIA A100-SXM4-40GB を 8 台使用した。各モデルの諸元と、製作に要した時間は、表 4 の通り。これら 3 つのモデルを、以下の HuggingFace Hub で公開した。

`https://huggingface.co/KoichiYasuoka/modernbert-small-classical-chinese-traditional`

`https://huggingface.co/KoichiYasuoka/modernbert-base-classical-chinese-traditional`

`https://huggingface.co/KoichiYasuoka/modernbert-large-classical-chinese-traditional`

さらに、esupar 1.8.2 の異体字テーブルを用いて [12]、各モデルの単語ベクトルを、中国の簡化字や日本の常用漢字に拡張した。この結果、語彙数 (vocab size) は 25078 とな

表 4: 古典中国語 ModernBERT 繁體字版の諸元

	small	base	large
hidden size	256	768	1024
intermediate size	768	1152	2624
local attention	128	128	128
max position embeddings	8192	8192	8192
num attention heads	8	12	16
num hidden layers	16	22	28
vocab size	23798	23798	23798
総パラメータ数	19M	123M	351M
製作に要した時間	7h8m	14h2m	24h6m

り、総パラメータ数は、small が 19M、base が 124M、large が 352M となった。これら 3 つのモデルを、以下の HuggingFace Hub で公開しつつ、品詞付与・係り受け解析ファインチューニングで用いることにした。

`https://huggingface.co/KoichiYasuoka/modernbert-small-classical-chinese`
`https://huggingface.co/KoichiYasuoka/modernbert-base-classical-chinese`
`https://huggingface.co/KoichiYasuoka/modernbert-large-classical-chinese`

なお、いずれのモデルも、製作用 python スクリプトを `maker.py` として含めておいたので、参考にされたい。

4 ModernBERT を用いた品詞付与・係り受け解析

古典中国語 ModernBERT を用いて、品詞付与と係り受け解析を同時におこなうアルゴリズムを考えてみよう。基本的手法としては、UD の各リンクに対する隣接行列に、UPOS と DEPREL の両方を埋め込む形で解析をおこなうことにする。

4.1 正方行列を用いた解析アルゴリズム

正方行列を用いた解析アルゴリズムでは、たとえば図 1 の UD 有向グラフに対し、単文字トークナイザに合わせるべく goeswith による変形 [3] をおこない (図 2)、以下の 6×6

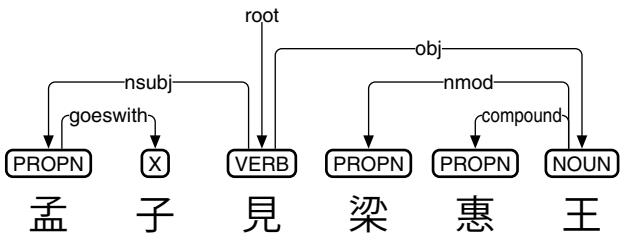
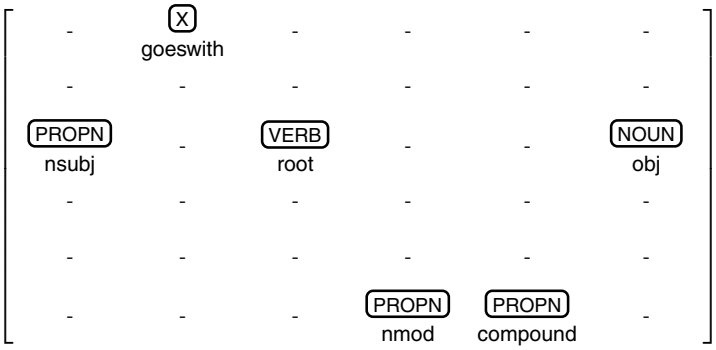


図 2: goeswith による図 1 の変形 (単文字トークナイザの場合)

の正方隣接行列を用いる。

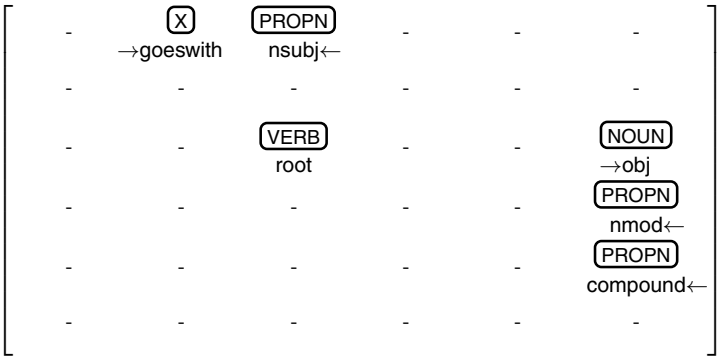


この正方隣接行列を、対数オッズ (logits) による確率行列の形で、系列ラベリングモデル上に実装する (図 3)。入力側では、各行の境目に [SEP] トークンを挟みこむと同時に、どの行に着目しているかがわかるよう [MASK] する。出力側では、正方隣接行列の各行を一次元に展開し、各行の境目は 1 トークンあけておく。正方行列を用いた解析アルゴリズムでは、UD 有向グラフのノード数 n に対し、入出力幅 $n(n+1)$ トークンの系列ラベリングモデルが必要[b] となる。

図 3: 正方行列を用いた系列ラベリング

4.2 上三角行列を用いた解析アルゴリズム

次に、上記の正方隣接行列を、上三角行列へと変換する。具体的には、DEPREL 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。



この上三角行列を、対数オッズ (logits) による確率行列の形で、系列ラベリングモデル上に実装する (図 4)。入力側では、各行の境目に [SEP] トークンを挟みこむ。出力側では、上三

[b] 入出力幅 8192 トークンの ModernBertForTokenClassification であれば、$n \leq 90$ の正方行列を乗せることができる。

角行列の各行を一次元に展開し、各行の境目は 1 トークンあけておく。上三角行列を用いた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[c]となる。

図 4: 上三角行列を用いた系列ラベリング

5 評価と考察

古典中国語 ModernBERT モデル (異体字拡張版 small・base・large) に対し、前章の 2 つのアルゴリズムによるファインチューニングを UD_Classical_Chinese-Kyoto [d]でおこない、評価 (lzh_kyoto-ud-dev.conllu による evaluation)・テスト (lzh_kyoto-ud-test.conllu による predict) をおこなった。評価指標は、CoNLL 2018 [13] の LAS (Labeled Attachment Score) / MLAS (Morphology-aware Labeled Attachment Score) / BLEX (Bi-LEXical dependency score) を用いた。評価結果を、表 5 に示す。

既存の古典中国語係り受けシステム (esupar による Biaffine [14] 実装) と比較してみたところ、正方行列アルゴリズムも上三角行列アルゴリズムも、既存システムに追いついていない。既存システムは、われわれが心血を注いでチューニングしてきたものなので、そう簡単に抜けるわけがないのだが、それにしても離され過ぎている。本稿の古典中国語 ModernBERT も、解析アルゴリズムも、まだまだ改良が必要だということである。

表 5: ModernBERT による古典中国語係り受け解析の評価 (LAS / MLAS / BLEX)

	評価 (evaluation)	テスト (predict)
modernbert-small-classical-chinese-ud-square	73.38 / 69.96 / 71.69	76.44 / 73.21 / 74.68
modernbert-base-classical-chinese-ud-square	74.26 / 70.97 / 72.48	77.47 / 74.30 / 75.78
modernbert-large-classical-chinese-ud-square	73.74 / 70.42 / 71.90	77.01 / 73.77 / 75.26
modernbert-small-classical-chinese-ud-triangular	73.41 / 70.24 / 71.72	76.64 / 73.36 / 74.94
modernbert-base-classical-chinese-ud-triangular	74.28 / 71.05 / 72.53	77.13 / 73.84 / 75.40
modernbert-large-classical-chinese-ud-triangular	72.88 / 69.44 / 71.14	76.32 / 73.11 / 74.63
既存システム (esupar 1.8.2)	82.13 / 78.29 / 79.22	82.06 / 78.21 / 79.27

[c]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。

[d]`https://github.com/UniversalDependencies/UD_Classical_Chinese-Kyoto`

参考文献

[1] 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, 藤田一乗: 古典中国語（漢文）Universal Dependencies とその応用, 情報処理学会論文誌, Vol.63, No.2 (2022 年 2 月), pp.355-363.

[2] 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.

[3] 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.

[4] 安岡孝一: GPT 系言語モデルによる国語研長単位係り受け解析, 人文科学とコンピュータシンポジウム「じんもんこん 2024」論文集 (2024 年 12 月), pp.83-90.

[5] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman: Universal Dependencies, Computational Linguistics, Vol.47, No.2 (June 2021), pp.255-308.

[6] Lucien Tesnière: Éléments de Syntaxe Structurale, Paris: C. Klincksieck (1959).

[7] Igor A. Mel'čuk: Dependency Syntax: Theory and Practice, New York: State University of New York Press (1988).

[8] Joakim Nivre: Towards a Universal Grammar for Natural Language Processing, CICLing 2015: 16th International Conference on Intelligent Text Processing and Computational Linguistics (April 2015), pp.3-16.

[9] 安岡孝一: Universal Dependencies にもとづく多言語係り受け可視化ツール deplacy, 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 (2020 年 12 月), pp.95-100.

[10] ウィッテルン・クリスティアン: 漢籍リポジトリ, センター研究年報 2015, 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 (2016 年 3 月).

[11] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush: Transformers: State-of-the-Art Natural Language Processing, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (October 2020): System Demonstrations, pp.38-45.

[12] 安岡孝一: Universal Dependencies と BERT/RoBERTa/DeBERTa/GPT モデルによる多言語情報処理 (2025 年 3 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 (2025 年 3 月).

[13] Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov: CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Proceedings of the CoNLL 2018 Shared Task (October 2018), pp.1-21.

[14] Timothy Dozat, Christopher D. Manning: Deep Biaffine Attention for Neural Dependency Parsing, 5th International Conference on Learning Representations (April 2017), C25.