日本語長単位語における語彙素推定

尾崎 太亮 [1] 古宮 嘉那子 [1] 浅原 正幸 [2] 小木曽 智信 [2]
[1] 東京農工大学大学院 生物システム応用科学府
[2] 国立国語研究所・総合研究大学院大学
hiroaki-ozaki@st.go.tuat.ac.jp, kkomiya@go.tuat.ac.jp,
{masayu-a, togiso}@ninjal.ac.jp

概要

デジタル人文学やコーパス言語学研究に重要な日本語長単位語に対する語彙素を推定する試みを行った．語彙素の推定にはエンコーダ・デコーダ型の言語モデル (T5) を用い，長単位語の品詞や表層形の他，長単位語を構成する短単位語の品詞・語彙素，および本文を入力として，長単位語の語彙素を出力させた．短単位語彙素をベースラインとした推定に対し，本手法は高い性能が得られた．検討の結果，短単位語は長単位語彙素推定に有効であるが，事例検討から，特に口語的な表現や動詞の付属部などに関しては文脈情報 (本文) が推定に有効であった．

1 はじめに

国立国語研究所が定義する日本語の語の認定単位の一つである長単位語は，文節認定ののちに文節を自立語部分と付属語部分の最大二単位に分割して得られる語の認定単位である．長単位語は，例えば，長大な固有名詞なども一語として認定されるなど，統計によるコーパス言語学やデジタル人文学の分野での解析・研究において利用価値が高い語の単位である．日本語はまた，形態素リッチな言語であり，かな・漢字などの複数の文字体系が混在している世界的にも珍しい特徴を有しているため，長単位語の語彙素の推定 (原形推定，Lemmatize) は語の統計に基づく言語学的な研究において，語の同定に必要な情報であり重要性が高い．

長単位語はその認定時の特性上，辞書の作成・管理が非常に困難である．これは，国立国語研究所が定義する短単位語とは対照的であり，短単位語においては UniDic[1] として形態素解析器 MeCab などで扱える形で辞書が提供されている[1)]．したがって，長単位の抽出や長単位の語彙素推定を高精度に実施するには，機械学習的手段を用いる必要性がある．

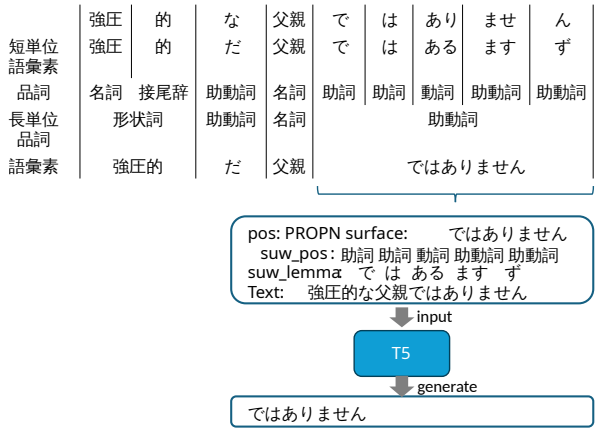

図 1 語彙素推定の概要．入力された短単位・長単位情報に基づき，長単位の語彙素を T5 モデルで推定する．

これまで尾崎らは，長単位語のスパン推定・品詞推定を実現する，文節・長単位解析器 Monaka の開発を行った [2]．Monaka は Transformer を用いた言語モデル (BERT) を用いることで，現代文から日本語の古文に至るまでの解析を可能とした．

本研究では，長単位語のスパン推定・品詞推定に加えて語彙素推定を実現することで，前述の言語学的な意義を踏まえ，日本語の特に現代語に対する長単位解析をひと揃え提供することを目的とする．

本研究においては，日本語の書き言葉として日本語書き言葉均衡コーパス (BCCWJ) と話し言葉として日本語日常会話コーパス (CEJC) を対象とした (1 章)．図 1 に示すように，長単位語の語彙素の推定には長単位語の品詞・表層形と長単位を構成する短単位語の表層形・品詞・語彙素などを入力とし，エンコーダ・デコーダ型の生成モデル (T5) を用いて長単位語の語彙素を出力とするモデルの学習と評価を行った (5，6 章)．評価の結果，前述の長単位・短単位の各情報のほか、本文を入力することで短単位語彙素をそのまま用いた長単位語彙素推定のベース

1) https://clrd.ninjal.ac.jp/unidic/

ラインを超える精度を得た (6 章)．本報告の学習・評価コードは，語彙素推定機能を長単位・文節解析 Monaka[2] に組み込み，公開をした．また，Monaka の長単位語彙素推定を含めた文節・長単位解析機能を搭載した Web アプリを Web 茶まめ[3] [3] 上で公開する予定である．

2 長単位・短単位とその語彙素推定

短単位語は，意味を持つ最小の単位を語種ごとに規定した上で，その最小単位を文節の範囲内で短単位の認定規定に基づいて結合させる（又は結合させない）ことにより，短単位を認定する．一方，長単位は文節を基にした単位であり，長単位の認定は，文節の認定を行った上で，各文節の内部を規則に従って自立語部分と付属語部分に分割していくことで行われる．この際，文節は一つまたは複数の短単位語で構成され，文節の分割も短単位境界を踏襲するため，長単位の境界は必ず短単位の境界となっており，長単位語は一つまたは複数の短単位語で構成される．

短単位語は UniDic[1] が整備されているため，MeCab[4] などの形態素解析器を用いることで，その語彙素を推定することができる．

一方長単位は，例えば，通常の名詞などは短単位語でもあり長単位語でもある場合がほとんどであるが，複合語や長い固有名詞，複数の助詞や助動詞からなる動詞の付属部 (例えば図 1) は，複数の短単位で構成される一つの長単位語として認定されることがほとんどである．そのため長い固有名詞を考えれば，その語彙素は正式な固有表現に根差した表記であり，その推定には単なる表記の正規化のみならず，エンティティリンキングに近い技術要素を含むタスクである．また長単位の認定が文節に依存しているため，長単位はその異なり語数が短単位に比べて膨大であり，前述の事情を含めて，その語彙素の推定には機械学習的な手段を用いる必要がある．

3 コーパス

日本語書き言葉均衡コーパス (BCCWJ) BCCWJ[5] [4] は書籍全般，雑誌全般，新聞，白書，ブログ，ネット掲示板，教科書，法律などのジャンルにまたがって 1 億 430 万語のデータを格納した均衡コーパスである．短単位・長単位のアノテーションが付与されている．

日本語日常会話コーパス (CEJC) CEJC[6] [5] は，日常場面の中で当事者たち自身の動機や目的によって自然に生じる会話を対象とし，多様な場面の会話をバランスよく集めたことを特徴とするコーパスである．CEJC に対し，その文節と長単位形態論情報，係り受け情報が付与されたデータ [6] が存在する．本研究においては，この長単位情報を用いた．

コーパス統計 同一の品詞・語彙素・同一の表層形の長単位語をユニークエントリとして，品詞ごとの異なり数，短単位の語彙素と一致する割合，および短単位と完全に一致する割合 (長単位と短単位が同一スパンでかつ同じ語彙素を持つ) を示したのが，表 1 である．短単位の語彙素との一致する割合から，短単位の語彙素を用いた長単位語彙素の推定のベースラインが 81〜84%程度であると言える．BCCWJ と CEJC ともに名詞と動詞が大半をしめるが，BCCWJ においては短単位と完全に一致する割合が低い傾向にある．また，BCCWJ と CEJC と大きく異なるのは，動詞，形容詞，感動詞，代名詞で，短単位の語彙素と一致する割合，完全に一致する割合ともに BCCWJ の方が低い．これらは図 1 のような文末表現・動詞の付属部ではないため，書き言葉の方が複合語を用いるケースが多いことが考えられる．副詞に関しては BCCWJ と CEJC において短単位の語彙素と一致する割合はともに高いが，CEJC においては短単位とほぼ一致していることが窺える一方，BCCWJ では短単位と完全一致する割合が半分程度である．書き言葉においては複雑な副詞的表現が，話し言葉よりも多いと言える．なお，表 1 にはすべての品詞の統計を記載しているわけではないが，CEJC においては「言いよどみ」が異なり数で 613 件あり，BCCWJ の 12 件に比べると圧倒的に多く，話し言葉の特徴が如実に現れている．CEJC の「言いよどみ」はすべて短単位と完全に一致していた．

4 関連研究

日本語における長単位に対する語彙素推定については筆者の知るところ，既存研究が存在しない．一方，長単位語彙素推定とは詳細な点で異なるが，Lemmatization という観点においては，Universal Dependencies (UD) [7] において Lemmatization が対象

2) https://github.com/komiya-lab/monaka
3) https://chamame.ninjal.ac.jp/
4) https://taku910.github.io/mecab/
5) https://clrd.ninjal.ac.jp/bccwj/
6) https://www2.ninjal.ac.jp/conversation/cejc.html

表 1 コーパス統計．品詞ごとの異なり数，短単位の語彙素と一致する割合，および短単位と完全に一致する割合を BCCWJ と CEJC に対して示す．

品詞/対象	BCCWJ			CEJC		
	異なり数	短単位語彙素	短単位	異なり数	短単位語彙素	短単位
全体	106,485	81.07	30.57	14,581	84.38	58.94
名詞	81,677	86.16	23.78	8,355	86.06	46.98
動詞	16,084	53.52	49.15	2,899	71.02	67.85
形状詞	2,450	93.76	54.53	504	95.24	65.87
副詞	2,203	91.60	55.11	651	99.08	90.78
形容詞	1,463	74.30	63.98	502	92.43	76.29
助動詞	676	42.31	37.86	330	50.91	45.45
助詞	350	71.43	66.29	120	78.33	75.00
感動詞	267	59.55	59.55	257	97.28	96.89
代名詞	186	79.57	55.38	92	100.0	77.17
接続詞	91	74.73	42.86	55	85.45	41.82

タスクに含まれ，例えば Stanza[7] [8, 9] などが UD における長単位を対象とした係り受けデータセット UD_Japanese-GSDLUW[8] を用いたモデルを提供している．これにおける Lemma は，そもそも日本語 UD において UniDic(短単位語辞書) と異なる語彙素を認定しているため，いわゆる長単位語彙素ではない[9]．なお，Stanza における Lemmatization の Pipeline はルールベースなどの適用すべきアルゴリズムを機械学習によって判別するなど複雑である．そのうちルール等で処理できない Lemmataization には seq2seq モデルを用いている [8]．Transformer を用いた言語モデルによる Lemmatization としては ByT5[10] を用いたものが研究が存在し [11]，通常の T5(mT5) に比べて，高性能であったとの報告がある．

5 語彙素推定方法

推定方法の概要を図 1 に示す．本研究においては，長単位のスパンと品詞は与えられるものとし，その上で長単位の語彙素を推定する．図 1 の例では，「強圧的な父親ではありません」という表現のうち，「ではありません」が一つの長単位語 (助動詞) となるが，その中には複数の動詞・助詞・助動詞である短単位語が含まれる．したがって，長単位表層形 (ではありません) と品詞 (助動詞) とそれぞれの短単位の語彙素・品詞，および本文を図 1 のように「pos:」などと情報の種別とともに空白区切りで一つの入力文字列とした．これを訓練済みエンコーダ・デコーダ型言語モデルによって語彙素を生成する．

データセット作成　まず学習データの作成には，各コーパスから同一の品詞・語彙素かつ同一の表層形をもつ長単位語例の抽出を行った．同一の品詞・語彙素・表層形の長単位語をユニークエントリとして，各コーパスごとにそれらのユニークエントリに対する諸要素 (品詞・表層形・語彙素・本文) と長単位を構成する短単位語の各要素 (品詞・表層形・語彙素) を抽出したものをデータセットとした．なお，各ユニークエントリは，コーパス中に複数登場することがあり，本文は各ユニークエントリに対して複数存在するが，最初に登場した本文のみを収録した．

上記の手順で各コーパスに対してデータセットを作成した後，そのデータセットから，学習 (train)・開発 (dev)・評価 (test) の組みを 3 組抽出した．各組ごとにデータセットのデータは全て用い，各組の評価セットは互いにデータが重複しないように抽出した．また，train/dev/test の比率は 90/5/5 である．

モデルとハイパーパラメタ　学習に使用したモデルは sonoisa/t5-base-japanese[10] である．学習率は 2e-5，バッチサイズは 24，データセットのサイズによらず学習ステップ数を 8000 とした．その他のハイパーパラメタは Transformers の Seq2SeqTrainer のデフォルト値[11] を用いた．

語彙素推定に対する各要素の影響を調査するため，以下の 4 つの場合を検討した．

- **Full**: 長単位品詞・表層形・本文と短単位品詞・表層形・語彙素を入力とするモデル．
- **w/o Text**: 長単位品詞・表層形と短単位品詞・表

7) https://stanfordnlp.github.io/stanza/
8) https://github.com/UniversalDependencies/UD_Japanese-GSDLUW
9) コーパス開発者らの独自の形態素解析器における語彙素認定基準を用いている．
10) https://huggingface.co/sonoisa/t5-base-japanese
11) https://huggingface.co/docs/transformers/v4.47.1/en/main_classes/trainer#transformers.Seq2SeqTrainingArguments

表 2　語彙素推定精度 (**Base** は表 1 の短単位語彙素との一致率，下線は **Base** を越えた結果，太字は最大性能である．)

品詞/対象	BCCWJ					CEJC				
	Full	**w/o Text**	**w/o SUW**	**LUW**	**Base**	**Full**	**w/o Text**	**w/o SUW**	**LUW**	**Base**
全体	80.67	**90.98**	68.69	79.33	81.07	**90.35**	85.41	77.59	56.20	84.38
名詞	79.47	**91.32**	70.88	82.64	86.16	91.58	85.32	85.65	64.39	86.06
動詞	86.22	**93.27**	66.16	74.58	53.52	**93.74**	92.84	71.81	46.53	71.02
形状詞	86.05	**95.35**	73.90	82.43	93.76	94.12	94.12	83.82	47.06	**95.24**
副詞	86.23	90.36	53.72	58.68	**91.60**	95.00	96.00	71.00	48.00	**99.08**
形容詞	85.02	**88.66**	48.58	53.44	74.30	91.14	**96.20**	55.70	32.91	92.43
助動詞	**80.00**	49.00	20.00	23.00	42.31	**75.47**	69.81	30.19	16.98	50.91
助詞	70.00	**78.33**	48.33	53.33	71.43	84.21	84.21	**89.47**	21.05	78.33
感動詞	56.52	54.35	28.26	23.91	**59.55**	88.24	94.12	32.35	8.82	**97.28**
代名詞	**84.62**	76.92	30.77	26.92	79.57	**100.0**	76.92	15.38	0.00	100.0
接続詞	72.73	45.45	63.64	54.55	**74.73**	28.57	57.14	42.86	28.57	**85.45**

層形・語彙素を入力とするモデル．

- **w/o SUW**: 長単位品詞・表層形・本文を入力とし，短単位情報を用いないモデル．
- **LUW**: 長単位品詞・表層形のみを入力とし，長単位情報のみを用いるモデル．

評価方法　評価は，まず各組の train セットで学習を実施し，評価セットで性能評価した結果を平均して算出した．評価に用いた指標は，正解の語彙素との完全一致の割合 (精度，accuracy) である．

6　評価と考察

評価結果　表 2 に評価結果を示す．まず，各ケースの比較から，短単位情報はすべての場合で有効であった．BCCWJ **w/o Text** と CEJC **Full** がそれぞれ最もよく，いずれも短単位の語彙素をそのまま用いるベースラインより高い性能を得た．品詞ごとに見れば，副詞・感動詞・接続詞は短単位語彙素をそのまま用いることが最も性能が高く，他は本手法による品詞推定が有効であった．特に動詞・形容詞・助動詞などの活用を伴う品詞において，深層学習の恩恵を得られやすい傾向が見られた．名詞において，深層学習によって精度が向上するのは主に固有名詞の影響だと考えられる．例えば，BCCWJ **w/o Text** において，名詞の短単位語彙素とも推定結果とも異なる結果となったエラーケース 148 件中 147 件が固有名詞で，残り 1 件が数詞であった．この数詞はアラビア数字で記載された電話番号であり，その語彙素はアラビア数字を漢数字に置き換えたものであった．

文脈情報 (本文) の有効性　BCCWJ において，本文を入力すると精度が下がる (**Full** と **w/o Text**，**w/o SUW** と **LUW** の対比) が，CEJC においては，その逆の傾向となった．品詞ごとに見ると，名詞・動詞において前述の傾向が見られたが，助動詞・代名詞においては一貫して文脈情報を必要とすることが窺える．例えば，助動詞に関しては図 1 にあるように，動詞付属部が一つの助動詞となる場合が多く，このことが関連していると考えられ，大半のエラーケースが複数の短単位語で構成される場合であった．代名詞に関しては「アタイら」(語彙素: あたい等) を「私達」などと推定する場合が見られた．これらの口語的な表現の推定精度の向上に文脈情報が寄与すると考えられる．

口語的表現の影響　口語的表現が多く出現する品詞として代名詞の他に接続詞が挙げられる．接続詞において，BCCWJ **Full** や **w/o Text** における，前述の名詞における調査と同様の調査をすると，表層形「しっかし」に対し，長単位語彙素は「しっかし」，短単位語彙素と推定結果はともに「然し」となるケースや，表層形「だったら」に対し長単位語彙素「だったら」であり，短単位語彙素は「だた」(2 語)，推定結果は「だった」である場合など，口語的な接続詞が多く見受けられた．CEJC においても当然，このような口語的な接続詞が数多くみられた．

7　おわりに

本研究では，長単位語の語彙素を推定を T5 を用いて行うことを検討した．検討の結果，短単位語彙素をベースラインとした推定に対し，高い性能が得られた．品詞ごとの推定結果では，短単位語彙素が最も精度が高い場合もあり，品詞ごとの有効性が大きく異なっていた．また，口語的な接続詞や動詞付属部の長単位の語彙素推定は難易度が高く，さらなる精度向上に向けた検討の余地があると考える．今後は，品詞ごとに取り扱いを変えるなど精度向上と，一般の公開を行う予定である．

謝辞

本研究は JSPS 科研費 JP22K12145，国語研「開かれた共同構築環境による通時コーパスの拡張」「多様な語義資源を統合した研究活用基盤の共創」「アノテーションデータを用いた実証的計算心理言語学」の助成を受けたものです.

参考文献

[1] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源 : 形態素解析用電子化辞書の開発とその応用. 日本語科学, 2007.

[2] Hiroaki Ozaki, Kanako Komiya, Masayuki Asahara, and Toshinobu Ogiso. Long unit word tokenization and bunsetsu segmentation of historical Japanese. In John Pavlopoulos, Thea Sommerschield, Yannis Assael, Shai Gordin, Kyunghyun Cho, Marco Passarotti, Rachele Sprugnoli, Yudong Liu, Bin Li, and Adam Anderson, editors, **Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)**, pp. 48–55, Hybrid in Bangkok, Thailand and online, August 2024. Association for Computational Linguistics.

[3] Toshinobu Ogiso and Tomoaki Tsutsumi. Webchamame: An online tool for morphological analysis of various historical japanese texts using unidic dictionaries. In Anne Baillot, Toma Tasovac, Walter Scholger, and Georg Vogeler, editors, **DH**, 2023.

[4] 前川喜久雄（監修）, 山崎誠（編）. 書き言葉コーパス—設計と構築—. 朝倉書店, 2014.

[5] 小磯花絵, 天谷晴香, 居關友里子, 臼田泰如, 柏野和佳子, 川端良子, 田中弥生, 伝康晴, 西川賢哉, 友香渡邊. 『日本語日常会話コーパス』設計と構築. 国立国語研究所論集, 2023.

[6] 浅原正幸, 若狭絢. 『日本語日常会話コーパス』に対する係り受け情報アノテーション. 言語処理学会第 28 回年次大会, pp. 1699–1703, 2022.

[7] 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. In Daniel Zeman and Jan Hajič, editors, **Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**, pp. 1–21, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[8] Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. Universal Dependency parsing from scratch. In Daniel Zeman and Jan Haji˘ c, editors, **Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**, pp. 160–170, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[9] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, 2020.

[10] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models, 2021.

[11] Krzysztof Wróbel and Krzysztof Nowak. Transformer-based part-of-speech tagging and lemmatization for Latin. In Rachele Sprugnoli and Marco Passarotti, editors, **Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages**, pp. 193–197, Marseille, France, June 2022. European Language Resources Association.