Monaka による BCCWJ2 における長単位情報の付与支援

尾崎 太亮 [1] 浅田 宗磨 [2] 古宮 嘉那子 [1] 近藤 明日子 [2] 小木曽 智信 [2,3]
東京農工大学大学院 生物システム応用科学府 [1] 国立国語研究所 [2] 総合研究大学院大学 [3]
hiroaki-ozaki@st.go.tuat.ac.jp asada@ninjal.ac.jp
kkomiya@go.tuat.ac.jp kondo@ninjal.ac.jp togiso@ninjal.ac.jp

概要

現代日本語書き言葉均衡コーパス (BCCWJ) の拡張 (BCCWJ2) に向けて，人手アノテーション (コアデータ) と非コアデータ作成のために，深層学習による長単位情報付与について検討を行った結果について報告する．従前の BCCWJ を対象とした検討で，正解率 99.80%，BCCWJ2 において 99.45%という実用的な性能を実現した．また異なる事前学習済み言語モデルを比較から形態論情報についてはスケール則が単純に当てはまらないことが示唆された．さらに，長単位付与のエラーからは，長単位の認定基準の更新につながる示唆が得られた．

1 はじめに

2024 年度から開始された「信頼できる言語資源としての現代日本語の保存・活用のためのデジタル基盤整備事業」において現代日本語書き言葉均衡コーパス (BCCWJ) の拡張を行っている．この拡張を BCCWJ2 と呼ぶ．この拡張の背景としては，これまでの BCCWJ が 10 年以上前のデータであるためより新しいデータが望まれることと，コーパスの規模自体の拡充の必要性がある [1]．BCCWJ2 では効率的に短期間で集中してコーパスを整備していく必要から，簡素化やアウトソーシングなどが行われているが，効率的なコアデータの作成と，精度の高い非コアデータの提供の観点から，効率的な形態論情報の付与も望まれる．

国立国語研究所 (以下，国語研) が定義する形態論情報は，短単位語と長単位語の二種類からなるが，短単位に関しては UniDic 辞書 [2] を用いた MeCab によって高精度での解析が可能である．BCCWJ2 の作成においては，コアデータに対して UniDic 辞書のエントリー追加も行われる見込みであり，これにより非コアデータの解析精度向上も期待されるところである．一方，長単位においては，BCCWJ 作成時には Comainu が用いられていたところ，BCCWJ2 においては深層学習ベースの長単位解析器である Monaka[1)] を用いることとなった．そこで，BCCWJ のコアデータを学習・評価データとした詳細なモデルチューニングを行った．モデルチューニングにおいては，特に昨今の日本語で学習された事前学習済み言語モデルの精度等を比較した．本稿では，これらの結果と，BCCWJ2 における評価について述べる．

表 1 BCCWJ のレジスターと語数 (**太字**はコアデータあり，語数の括弧内はコアデータでの語数)

サブコーパス	レジスター	語数 (万)
出版	**書籍 (PB)**	2,855 (23)
	雑誌 (PM)	444 (36)
	新聞 (PN)	137 (24)
図書館	書籍 (LB)	3,038
特定目的	**白書 (OW)**	488 (23)
	教科書 (OT)	93
	広報紙 (OP)	376
	ベストセラー (OB)	374
	Yahoo!知恵袋 (OC)	1,026 (11)
	Yahoo!ブログ (OY)	1,019 (12)
	韻文 (OV)	23
	法律 (OL)	108
	国会会議録 (OM)	510

2 BCCWJ と BCCWJ2

表 1 には BCCWJ における各レジスターごとの語数を示す．図書館における書籍 (LB) が 3,000 万語以上と最大となっているが，コアデータを含まない．出版サブコーパスは全てのレジスタでコアデータをもち，同様に書籍 (PB) が最も大きく LB とほぼ同等の規模となっている．他に，白書 (OW) や，Yahoo!知恵袋 (OC)，Yahoo!ブログ (OY) においてコアデー

1) https://github.com/komiya-lab/monaka/tree/dev/bccwj2
なお，Monaka は長単位のほか文節や文節係り受けの解析も可能である．

タが存在し，OC と OY は 1,000 万語以上と特定目的サブコーパスにおいて最も大きなレジスターとなっている．

コアデータは各 11〜36 万語と大きな違いはなく，コーパス全体のサンプリング比率とはやや異なっている．コアデータ全体としては 129 万語である．

BCCWJ2 においては，全てのデータが出版サブコーパスの書籍 (PB) となる．これまでと同様に，国立国会図書館提供の書誌のうち，外国語で書かれた書籍のほか，マンガ・絵本，学習参考書・試験問題集，要覧・名簿・電話番号簿，写真集・月報・小冊子・大型本・豆本，官庁資料，外国の教科書，脚本，手稿，書誌・目録，百科事典・一般年鑑，博士論文，地図，ピアノ譜等を除外したものが用いられている．

BCCWJ2 のデータは BCCWJ に未収録の 2006 年以降のデータが順次拡充されていく見込みである．本検討においては 2008 年のデータを用いた．2008 年のデータのうち，コアデータは短単位で約 19 万語，非コアデータが約 513 万語である．コアデータにはすでに人手修正済みの短単位情報が付与されている．

3 解析モデルの改良

3.1 品詞以外の形態論情報

長単位を構成する品詞以外の形態論情報として，書字形，語彙素，語彙素読みがある．長単位解析器 Monaka は短単位を入力として長単位境界と品詞情報を提供するものである．語彙素の推定も別途行うことが可能であるが，以前の BCCWJ 構築においては長単位のこれらの情報は，短単位情報から組み上げて作成されているため，語彙素推定は行わずに BCCWJ2 でもそのプロセスを継承した．

書字形や語彙素，語彙素読みは短単位のそれらを基本的には結合させたものを用いた．活用型や活用形は長単位を構成する短単位のうち，最後尾の用言が持つ活用型と活用形を用いた．

3.2 平均アンサンブルを用いた推論

後述する通り，モデル学習においては 3 分割交差検定を用いて評価を行ったため，各種類ごとに 3 つの学習済みモデルが存在する．BCCWJ2 に対してこれらを適用する際には，平均アンサンブルを用いた解析を行った．具体的には，モデルの最終層の出力 (ラベルと同数の次元をもつ出力) を平均化した上で，最大値を取るラベルを解析結果として採用した．

3.3 ルールによる解析後処理

長単位語の認定や品詞について，長単位語を構成する短単位語の様態からほぼ一意に定まるような状況に対しては，ルールによる解析後処理によって修正を行った．本検討で追加したルールは全て，長単位語が一つの短単位語のみで構成されるケースについてである．このケースにおいては，長単位品詞は短単位のものと同じものとなるが，「名詞-普通名詞-形状詞可能」，「名詞-普通名詞-サ変形状詞可能」，「名詞-普通名詞-副詞可能」の短単位品詞については，長単位品詞が名詞の他，形状詞または副詞となる可能性がある．このため，これらのケースについては解析器の出力する品詞に対して，該当の品詞のうち最もスコアが高いものを用いることとした．

4 評価方法

本稿では，解析器 Monaka の学習を Ozaki ら [3] にしたがって行った．ハイパーパラメタは付録 B の表 5 の通りである．学習・評価には，コアデータから一定比率で学習・検証・評価データを区分したデータを 3 組作成し，検証データで最も高い性能であったエポックの評価データにおける評価指標を用いた．学習・検証・評価データの比率は，80%，10%，10%となっている．このうち 3 つの評価データは，互いに重複が無いようにサンプリングした．

4.1 比較対象モデル

本研究では，ハイパーパラメタは同一で，言語モデルのみを変えて解析器の学習を行った．

- **東北大 BERT-base**[2] を用いたモデル．サブワード分割は IPA 辞書に基づく MeCab(fugashi) 単位を基礎として行っている．日本語 Wikipedia 約 3,000 万文と日本語 CC-100 約 3 億 9,200 万文で学習されたモデルである．最大トークン数は 512 である．モデルパラメタ数は約 111M である．
- **Megagon Labs RoBERTa** [3] 早大 RoBERTa[4] のトークナイザを用いたモデルであり，Juman++

2) tohoku-nlp/bert-base-japanese-whole-word-masking
3) megagonlabs/roberta-long-japanese
4) nlp-waseda/roberta-base-japanese

表 2 長単位境界 (書字形) の正解率

モデル	PB	PM	PN	OW	OC	OY	マクロ平均
東北大 BERT-base	99.71	99.28	99.52	99.83	99.29	0.047	83.72
Megagon Labs RoBERTa	99.69	99.63	99.55	99.90	99.52	99.32	99.60
TohokuNLP BERT-alpha 500M	99.58	99.48	99.46	99.86	99.28	99.07	99.45
日本語 BigBird	**99.80**	99.75	**99.75**	**99.96**	99.46	**99.63**	**99.72**
RetrievaBERT	**99.80**	**99.80**	99.71	99.95	**99.61**	99.42	99.71

を用いたサブワード分割を行う．Japanese mC4 (約 2 億文) で学習されたモデルである．モデルパラメタ数は約 111M である．最大トークン長は 1282 である．

- **TohokuNLP BERT-alpha 500M**[5] llm-jp-v3 tokenizer[6] を用いており，Whole Word Masking 単語分割に用いた辞書は UniDic に基づくものである．学習データは llm-jp-corpus-v3[7] の日本語コーパスのサブセット (ja_cc, ja_warp_html, ja_warp_pdf, ja_wiki, kaken) を用いたものである． パラメタ数は約 581M であり，最大トークン長が 8192 であるモデルを用いた．
- **日本語 BigBird**[8] Juman++を用いたサブワード分割を行うモデルであり，学習データは日本語 Wikipedia と日本語 CC-100，Japanese OSCAR である．BigBird[4] はロングコンテキスト向けのモデルであり，4096 トークンに対応している．モデルパラメタ数は約 113M である．
- **RetrievaBERT**[9] Japanese CommonCrawl, RefinedWeb, Chinese Wikipedia, Korean Wikipedia, The Stack で学習されたモデルである．パラメタ数は約 1.45B であり，2096 トークンまで入力できる．

BCCWJ2 の非コアデータでは，現時点において，サブワードで 512 トークンを超える文が存在するため，最大トークン数が大きいモデルを対象にモデル探索を行った．なお，東北大 BERT-base は最大 512 トークンまでであるが，Ozaki ら [3] の学習に用いられているため，比較のために利用した．この他，ModernBERT の利用も検討したが，現状上手く学習が行えない問題が発生したため，今後の課題としている．

5) tohoku-nlp/tohokunlp-bert-500m-sq8192-alpha
6) https://github.com/llm-jp/llm-jp-tokenizer
7) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3
8) nlp-waseda/bigbird-base-japanese
9) retrieva-jp/bert-1.3b

4.2 評価指標

BCCWJ においては，基本的に短単位を 1 レコード (行) として管理している．このため，このレコードごとに対する正解率を評価指標として用いた．

各行にはコーパスのメタ情報のほか，短単位の情報と，長単位の情報が記述される．この際，長単位が複数短単位で構成される場合は，先頭の短単位に対応する行にのみ，長単位先頭のフラグと共に長単位の各情報が記載される．先頭ではない短単位の行には長単位情報は何も記載されない．

このことから長単位の境界推定の指標としては，長単位書字形の正解率を用いた．これは，書字形は長単位境界が正確に一致する場合のみ，各行での書字形が一致するためである．また，長単位品詞の推定結果も併せて評価した．

BCCWJ2 のコアデータに対しては，2000 語をサンプリングし，人手で解析結果評価した上，主な解析エラーを明らかにした．また非コアデータに対しては処理時間を計測して実用性を評価した．なお，学習・解析に用いた環境は，RTX A6000 (48GB GPU メモリ) 376GB CPU メモリ，Intel(R) Xeon(R) Gold 6238R CPU 2.20GHz である．推論時のバッチサイズは長文でも GPU メモリの余裕がある 8 とした．

5 評価結果

5.1 解析精度

表 2 には，長単位境界推定の評価としての書字形における正解率を示す．OY(Yahoo!ブログ) を除き，どのモデルも 99%以上の精度を実現している．しかしながら，特にコアデータにおいては，長単位の推定の後，人手での修正を行うことから，修正要否の観点においてエラー率にも着目するべきである．例えば，PB レジスタにおいて最良のモデルは最も正解率が低いモデルの 1/2 以下のエラー率であり，人手修正を半分以下に減らすことができることを意味するため，正解率には明確な差があると見るべきで

ある．

BCCWJ2 の対象となっている PB(書籍) は，OW(白書) に次いで高い精度のレジスターとなっている．東北大 BERT-base の OY では，特異的に精度が低下しているが，これは，東北大 BERT-base では全角空白文字がエンコードされないため，行ずれが発生してしまったことによる．一方で，OY はどのモデルもあまり精度が高くなく，ブログ特有の記述や URL を含むような記載内容の幅広さ故の処理の難しさが窺える．

全般的に最も正解率が高いモデルは，日本語 BigBird であった．次いで，ほぼ同様の正解率となったのは RetrievaBERT であった．BERT-alpha 500M や RetrievaBERT は日本語 BigBird に比べて，5 から 10 倍程度大きな言語モデルであるので，長単位解析においてはいわゆるスケーリング則のような傾向は見られないと考えられる．この点は，言語モデルにおいて低レイヤーで獲得されるような基盤的なタスクについては，日本語 BigBird 程度のモデルサイズで十分知識を獲得できる可能性が示唆される．また，BigBird モデルは長文入力に対応するために注意機構を「間引く」ような処理が入っている．このため，注意機構によって文脈中の他のトークンから情報を引き出せる他のモデルとは異なり，BigBird の推論能力の向上の過程として単体のトークンレベルでの埋め込み表現の「質」が向上していることも考えうる．

これらの結果については長単位品詞についても同様であった．正解率のマクロ平均が最も高いのは RetrievaBERT であったが，PB での正解率が高かったのは日本語 BigBird であった．詳細については，付録 A に記載した．

5.2 BCCWJ2 における評価

表 3 BCCWJ2 での評価

		誤り数	正解率
空白・補助記号を含む	境界	11	99.45
	品詞	24	98.80
空白・補助記号なし	境界	12	99.40
	品詞	26	98.70

BCCWJ2 における評価は Megagon Labs RoBERTa の結果を用いた．これは単に学習・評価を並行して進めている都合上，最初に実用的な性能を達成した当該モデルでの評価を行なったものである．

表 3 には BCCWJ2 での評価結果を示す．評価対象の 2000 語のサンプルに空白・補助記号を含む場合と含まない場合，それぞれについて評価した．長単位境界に関しては，BCCWJ での評価とほぼ同様の正解率であった．品詞推定についてはやや低い結果であるが，実用的に十分高い正解率であると言え，2008 年のデータにおいては従前の BCCWJ コアデータを用いたモデルで十分な性能を実現できた．

5.2.1 主な解析エラー

主な解析エラーに，感動詞に関するもの，記号に関するもの，並列的な動詞の運用に関するもの，外来語や文語などのドメイン外の入力に関するものがあった．このうち感動詞に関しては，「さあさあ」などの感動詞の繰り返しや感動詞+格助詞「と」(ハーッと) などの長単位 2 語が長単位 1 語と推定されるものであった．また並列的な動詞の運用例として，「伝え合い共有する」のような長単位 2 語 (伝え合い/共有する) が一語となるものがあった．このようなケースは，人手によっても長単位認定が難しいケースであると考えられ，より明確な規定の整備等も含めて検討する必要がある．

5.2.2 処理時間

BCCWJ2 の 2008 年における非コアデータを対象に処理時間を計測した．非コアデータは約 513 万語あり，日本語 BigBird で 45 分要した．今後ルールの追加や長単位認定基準の変更などでモデルや解析器に変更を加える可能性を考えても，十分実用的な時間で処理することができた．

6 おわりに

本稿では，現代日本語書き言葉均衡コーパス (BCCWJ) の拡張 (BCCWJ2) に向けて，形態論情報のうち特に長単位情報付与について言語処理の観点から精査を行った．この結果，書籍において正解率 99.80%という実用的な性能を実現した．また異なる事前学習済み言語モデルを比較したところ，長単位情報の付与には日本語 BigBird などの比較的小規模なモデルが最も良い性能を達成した．エラー解析からは，推論時のルール追加のほか，長単位の認定基準の更新につながる示唆が得られるなど，深層学習モデルと人文・国語学的な検討において相補的な取り組みの例としても有意義であったと考える．引き続き BCCWJ2 の更新に合わせて検討を継続する．

謝辞

本研究は文化庁委託事業「信頼できる言語資源としての現代日本語の保存・活用のためのデジタル基盤整備事業」の成果の一部です.

参考文献

[1] 山崎誠, 高橋雄太, 小木曽智信. 「現代日本語書き言葉均衡コーパス」の拡張 —bccwj2 の構築—. 言語処理学会第 31 回年次大会, pp. 414–417, 2025.

[2] 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.

[3] 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.

[4] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: transformers for longer sequences. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

表 4 長単位品詞の正解率

モデル	PB	PM	PN	OW	OC	OY	マクロ平均
東北大 BERT-base	99.61	99.25	99.47	99.82	99.17	9.94	84.55
Megagon Labs RoBERTa	99.56	99.31	99.29	99.84	99.32	98.96	99.38
TohokuNLP BERT-alpha 500M	99.33	99.18	99.24	99.82	98.92	98.50	99.16
日本語 BigBird	**99.73**	99.62	99.66	**99.92**	99.45	99.17	99.59
RetrievaBERT	99.71	**99.69**	**99.69**	99.92	**99.49**	**99.25**	**99.62**

A 長単位品詞の正解率

長単位の品詞推定においても長単位境界の正解率と同様の傾向を示した．境界推定との違いとしては，新聞レジスター (PN) と Yahoo!ブログ (OY) において日本語 BigBird と RetrievaBERT との正解率の順位が逆転している．このことは RetrievaBERT がより多様なドメインに対応していることが窺える．OY は境界推定よりも品詞推定の難易度が高いように見える．これは学習にない語 (外来語等) が多く登場し，それらの品詞を推定することが難しいと考えられる．

B ハイパーパラメタ

表 5 ハイパーパラメタ

パラメタ	値
dropout 率	0.5
POS embedding 次元数	256
学習率	5e-06
バッチサイズ	24
エポック数	20
gradient clip	5.0
gradient decay	0.75
decay step	5000