BERT による辞書推定システムを用いた 近代以前の日本語文書の形態素解析の精度向上

臼井久生 [1] 古宮嘉那子 [1] 小木曽智信 [2]
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 研究系 / 総合研究大学院大学 [2]
{h-usui@st., kkomiya@}go.tuat.ac.jp, togiso@ninjal.ac.jp

概要

本研究では，BERT により形態素解析用の辞書の推定を行うことで，近代以前の日本語文書の形態素解析の精度を向上させる手法を提案する．形態素解析用の辞書作成時に用いたデータを用いて BERT を fine-tuning し，文書分類を行うシステムを作成した．形態素解析の対象文書を，作成した文書分類システムに入力し，出力された文書クラスの辞書を用いて形態素解析を行った．また，二つのベースラインとその文書の正解の辞書を用いた手法（Oracle 手法）と比較した．実験の結果，提案手法はベースラインを常に上回り，いくつかの文書では Oracle 手法を上回った．

1 はじめに

日本には数多くの近代以前の日本語文書が存在している．未知の近代以前の日本語文書に対して形態素解析を行う際，専門家であっても文書がどの時代のもので，どのような文体の文書であるかを判断するのが難しい場合がある．このような場合において，通常は中古和文用の辞書を用いることが多い．これは中古，古典文学作品が成立した平安時代の古典文学が代表的なものであり，またその文体を継承して書かれた文献の量が多いためである．だが，このような場合において正しい辞書を用いることが出来れば，より正確な形態素解析が可能になる．本論文ではこのような場合に対し，文書分類システムを使用して適切な辞書を選択する方法を提案する．本研究では，BERT を使用した文書分類システムが文書の時代や文体を正しく特定できることを示し，推定された時代や文体に基づいた適切な辞書を使用することで，形態素解析の性能が大幅に向上することを実証した．

2 関連研究

間淵ら [1] は，異なる文体の混在するテキストに対する複数辞書の切り替えによる解析手法を提案している．彼らは，文体の異なるテキストに対して，複数の辞書を用いて形態素解析を行うことで，解析精度を向上させることができることを示している．しかし，彼らの手法は，文体の異なるテキストを事前に分類する必要があるため，未知の文書に対しては難しい手法である．

BERT を用いた多クラス分類は広く研究されており，上坂 [2] の研究や，林ら [3] による研究者の web ページの多クラス分類や，郷原 [4] らによる，BERT を用いた日本語文章の難易度の推定のための他クラス分類などがある．また，古文の UniDic 辞書は，小木曽ら [5, 6, 7, 8, 9, 10] によって作成されており，多くの時代や文体に対応するため多くの種類が存在している．

DNN を用いた形態素解析研究に，森田ら [11] による RNN を用いた形態素解析研究，植田ら [12] による汎用言語モデルによる日本語解析の一環としての形態素解析研究がなされている．これらの結果から，DNN を用いる場合，より精度が上がる事が示されているが，Kudo[13] が作成した MeCab との差は小さく，またより高速に形態素解析を行えるため，MeCab を用いることが一般的である．Web 上での形態素解析補助ツールに，堤ら [14] が作成した「Web 茶まめ」があり，多くの研究者に利用されている．しかしこのツールは辞書を推薦する機能はなく，ユーザーが手動で選択する必要がある．

3 BERT による辞書推定システム

図 1 に，提案手法の概要を示す．未知の近代以前の日本語文書に対して適切な辞書を選択するため

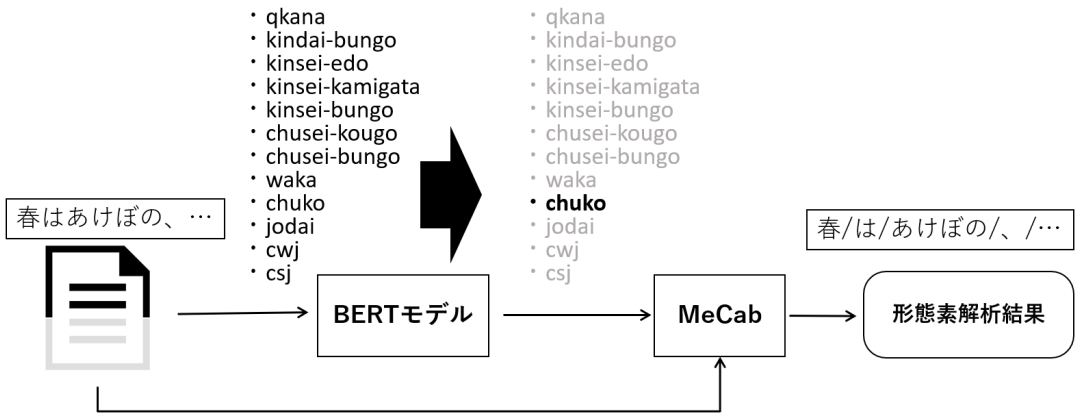

図 1　提案手法の概要図: 形態素解析したい文書の先頭 512 トークンを BERT に入力し，得られた辞書の推定結果と文書を MeCab に入力し形態素解析を行う．

に，BERT を用いた文書分類器を利用する．今回利用する BERT は入力テキストごとに最大 512 トークンを受け付けるため，分類したい文書の冒頭 512 トークンを抽出し，入力として使用した．出力は，入力文書の文書クラスである．この文書クラスの辞書を形態素解析に利用する．

4　UniDic データ

BERT の fine-tuning には，古文 UniDics の辞書作成時に用いたデータを使った．分析には年代やジャンル（物語，随筆，教科書など）が異なる以下の 12 種類の UniDic 辞書を利用した．

- Qkana: 旧仮名遣いによる口語体の日本語
- kindai-bungo: 近代（1870〜1945 年）における文語体の日本語
- kinsei-edo: 近世江戸（1603〜1869 年）における話し言葉の日本語
- kinsei-kamigata: 近世上方（京都・大坂）における話し言葉の日本語
- kinsei-bungo: 近世における文語体の日本語
- chusei-kougo: 中世（1185〜1602 年）における話し言葉の日本語
- chusei-bungo: 中世における文語体の日本語
- chuko: 中古（794〜1185 年）の和文体の日本語
- jodai: 上代（〜794 年）の日本語
- waka: 中古・中世の和歌
- CWJ[15][16]: 現代における文語の日本語
- CSJ[15][17]: 現代における話し言葉の日本語

今回 BERT の fine-tuning に用いたデータは，これらの辞書作成時に小木曽らが用いたデータのうち，その時代や文体が同じものを抽出して利用した．そのため，付録の表 3 に示すように，データ数には大きな差がある．

5　BERT の文書分類による辞書推定

5.1　実験

4 章で挙げた辞書に基づくコーパスを使用して 3 分割交差検定を実施した．この際，文書の作者が重複しないように辞書ジャンルごとに train，valid，test に分割した．例えば jodai に含まれるデータは，祝詞，続日本紀，萬葉集の 3 作であるため，それぞれの作品を train，valid，test に分割した．

表 3 で示すように，今回用いたコーパスの各ドメインの文書数は大きく差があり，最も少ない kinsei-edo で 46 件，最も多い CWJ で 2211 件であった．このため，文書分類の精度向上のために通常のクロスエントロピー損失関数に加え，クラスバランス型クロスエントロピーを使用した．通常のクロスエントロピー関数を式（1）で示す．

$$L_{CE} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{1}$$

ここで，N はクラス数，y_i は真のラベル，p_i は予測された確率である．クラスバランス型クロスエントロピー関数を式（2）で示す．

$$L_{BCE} = -\sum_{i=1}^{N} \frac{1-\beta}{1-\beta^{N_i}} y_i \log(p_i) \tag{2}$$

ここで，N_i はそれぞれの真のラベルのコーパスごとのデータ数，β は定数である．

今回用いたモデルは retrieva-jp/t5-base-long で，学習率は 7e-06，エポック数は 50，β は 0.99 であった．

5.2 評価

評価設定として，各クラスの 1-best と 3-best の正解率を計算した．1-best 正解率は，最も高い確率を持つクラスを選択した場合の正解率であり，3-best 正解率は上位 3 つのクラスの中に正解が含まれる場合を正解とした場合の正解率である．表 1 に，損失関数ごとの各クラスの 1-best と 3-best の正解率を示す．表中の太字の部分は，クラスバランス型クロスエントロピーを使用した際に正解率が上がったクラスである．また，この実験のベースラインは，通常のクロスエントロピーを使用した 1 -best 正解率である．

表 1 BERT による辞書推定の正解率 (%)．1-best，Normal がベースラインである．(太字：Normal と比較して Baranced の正解率が上昇した数値)

辞書名	1-best	1-best	3-best	3-best
Loss function	Normal	Balanced	Normal	Balanced
Qkana	68.12	59.81	72.30	**91.72**
kindai-bungo	68.70	67.04	88.30	79.38
kinsei-edo	11.11	**26.67**	68.80	**84.44**
kinsei-kamigata	33.82	**59.26**	80.00	**88.14**
kinsei-bungo	49.01	40.55	78.40	**79.70**
chusei-kougo	49.18	**69.34**	23.50	**53.11**
chusei-bungo	84.92	**92.75**	89.60	**97.39**
waka	51.88	42.02	72.80	**79.27**
chuko	53.35	44.25	83.80	81.22
jodai	30.95	30.28	36.50	**49.07**
cwj	95.45	92.41	88.90	**98.79**
csj	97.87	94.89	99.70	**99.86**
Ave,	57.86	**59.94**	73.55	**81.84**

結果として，通常のクロスエントロピーを使用したマクロ平均正解率は 57.86%であったが，クラスバランス型では 59.94%であった．3-best 正解率の評価でもクラスバランス型の方が高い結果を示した．

また，この時の混同行列を付録の図 2 に示す．この混同行列を見ると，いくつかのクラスでの混同が見られる．最も頻繁に見られる誤りは，chusei-bungo と chuko の混同であった．この理由として，中古期の文語体の文体が後の時代に引き継がれたためと考えられる．中世は中古期の後に続く時代であり，中世の文学には中古の文学の文体が含まれていることから，この混同が起きやすいと推察される．次に，waka と chuko の混同について注目する．これらは同時代に属しており，テキストの使用される文体が異なるのみであるため，解析上重大な問題とはならない．中古の文学の文中には多くの和歌が含まれているため，両者が混同されることは自然なことである．同様に Qkana，kindai-bungo の混同についても，ほぼ同時期に成立した文語体であるため，差異は小さい．また，CSJ と CWJ の混同について，この混同は，一部の人々が話し言葉のような文体で現代日本語を書き表すことがあるために起こると考えられる．総じて，これらの誤りは日本語学の研究者にとっても理解可能であり，妥当なものといえる．

6 形態素解析による評価

6.1 評価設定

文書分類システムを用いて，未知の文書に適した辞書を推定し，形態素解析を行った．評価には以下の，辞書及び文書分類システムの学習に利用されていない 4 種類のコーパスを使用した．

- 日本語読本 1（1936 年）：近代文語で書かれた，ハワイで用いられた日本語の教科書である．Kindai-bungo に該当する．
- 日本語読本 2（1936 年）：近代の話し言葉で書かれた，ハワイで用いられた日本語の教科書である．Qkana に該当する．
- 天草版金句集（1593 年）：中世文語で書かれた，キリスト教宣教師のための読本である．Chusei-bungo に該当する．
- 古今集遠鏡（1793 年）：本居宣長によって書かれた，古今和歌集の書評である．近世文語で書かれ，Chuko と kinsei-kamigata に該当する．

各テキストを 3 回分類し，3 分割交差検定に基づくマクロ平均 F 値を計算した．評価の比較には，Chuko 辞書のみを使用する Chuko ベースライン，CWJ 辞書のみを使用する CWJ ベースライン，および正しい辞書を使用する Oracle 手法を使用した．これらのベースラインについて，chuko は近代以前の日本語文書の内最も多い文体であるためこれを利用した．CWJ は全データのうち最も多い文体であるためこれを利用した．形態素解析の評価には以下の 4 つのレベルを適用した．

- Lv.1: 単語境界の正確さ
- Lv.2: 品詞や活用形の正確さ
- Lv.3: 語幹や発音の正確さ
- Lv.4: 発音の詳細（連濁などを含む）

この評価の際，1-best，クラスバランス型クロスエントロピーを使用した BERT モデルを利用した．

表 2　辞書推定による形態素解析のマクロ平均 F 値 (%): 提案手法が Oracle 手法を上回った場合にはアスタリスク（*）がついている．太字は Oracle 手法以外で最も高い数値

	レベル	日本語読本 1	日本語読本 2	天草版金句集	古今集遠鏡	平均
Chuko	Lv. 1	86.96	79.05	92.92	89.86	87.20
CWJ	Lv. 1	90.02	92.25	90.49	83.07	88.96
Proposed	Lv. 1	**95.80**	**96.57**	**96.06***	**92.43***	**95.21**
Oracle	Lv. 1	98.67	98.78	95.72	91.29	96.11
Chuko	Lv. 2	81.13	57.66	85.05	73.57	74.35
CWJ	Lv. 2	73.40	88.25	73.83	63.64	74.78
Proposed	Lv. 2	**89.93**	**91.72**	**90.41***	**77.80***	**87.47**
Oracle	Lv. 2	97.28	96.52	89.84	74.32	89.49
Chuko	Lv. 3	79.47	56.21	84.00	72.44	73.03
CWJ	Lv. 3	72.64	87.28	73.09	62.27	73.82
Proposed	Lv. 3	**88.98**	**91.04**	**89.79***	**76.67***	**86.62**
Oracle	Lv. 3	96.99	95.98	89.23	73.19	88.85
Chuko	Lv. 4	75.81	53.22	82.06	71.48	70.64
CWJ	Lv. 4	71.34	86.26	71.31	61.11	72.51
Proposed	Lv. 4	**86.83**	**90.03**	**87.99***	**75.89***	**85.18**
Oracle	Lv. 4	96.24	94.97	87.54	72.24	87.75

それぞれのデータの分類結果について付録の表 4 に示す．日本語読本 1 は Kindai-bungo, Kinsei-bungo, Kinsei-edo に分類された．日本語読本 2 は Qkana, Kindai-bungo，CWJ に分類され，天草版金句集は Chusei-bungo，Kinsei-bungo に分類された．古今集遠鏡は Kinsei-bungo および Chuko に分類された．

6.2　評価

表 2 に，形態素解析における各辞書のマクロ平均 F 値を示した．最良の結果は太字で示し，提案手法が Oracle 手法を上回った場合にはアスタリスク（*）がついている．

提案手法のマクロ平均 F 値は 4 つのテスト文書全体において評価のレベルにかかわらず 2 つのベースライン手法を上回った．最も右側の列は，4 つのテスト文書における平均 F 値を示しており，ベースラインと比較すると，レベル 1 評価では提案手法の F 値が 95.21%で最も高く，Chuko と CWJ の F 値がそれぞれ 87.20%と 88.96%であったことが確認できる．レベル 4 評価においても，提案手法の F 値は 85.18%で最も高く，Chuko と CWJ の F 値はそれぞれ 70.64%と 72.51%であった．

また，提案手法は天草版金句集と古今集遠鏡において，Oracle 手法を上回った．例を挙げると，レベル 1 評価において，天草版金句集に対する提案手法の F 値は 96.06%であり，Oracle 手法の 95.72%を上回った．

7　考察

コーパス全体の平均 F 値を比較した場合，Oracle 手法が提案手法より優れていた．しかし通常，Oracle 辞書を事前に知ることは不可能であるため，本実験においては提案手法が実用的に最も優れた手法であるといえる．本研究では形態素解析する際に，BERT による文書分類システムによって選択された辞書を使用した．その文書分類の結果について，文書分類システムは常に正解を導き出すことはできなかった．しかし形態素解析の結果においては，Oracle 手法よりも優れた場合があった．日本語学の研究者によると，この分類結果における分類ミスは形態素解析器にとって致命的な誤りではなく，類似した文書間の混同であると考えられる．以上の理由から，文書分類システムは形態素解析の精度向上に寄与したと結論づけられる．

8　まとめ

本研究では，適切な辞書を推定するための BERT 分類器を用いた文書分類システムを開発した．本システムの性能評価は，古文の形態素解析における F 値を指標とした．実験の結果，提案システムによって選択された辞書は，ベースラインを上回る性能を示したことが確認された．また，一部の文書においては，提案システムが選択した辞書が人手で選択された辞書よりも優れていることが明らかとなった．

以上の結果から，本研究で提案した文書分類システムは，古文解析において実用的かつ効果的な手法であることが示された．

謝辞

本研究は JSPS 科研費 JP22K12145，及び国立国語研究所共同研究プロジェクト「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」の助成を受けたものです.

参考文献

[1] 間淵洋子, 小木曽智信. 異なる文体の混在するテキストに対する複数辞書切り替えによる解析手法の提案. じんもんこん 2015 論文集, 第 2015 巻, pp. 125–130, dec 2015.

[2] 上阪彩香. Ca12-3 著者判別分析における形態素解析辞書選択. 日本行動計量学会大会抄録集 46, pp. 388–389. 日本行動計量学会, 2018.

[3] 林容央, 桂井麻里衣. 研究者の活動可視化に向けたウェブページの多クラス分類. 人工知能学会全国大会論文集 第 37 回 (2023), pp. 2P4GS1102–2P4GS1102. 一般社団法人 人工知能学会, 2023.

[4] 郷原聖士, 綱川隆司, 西田昌史, 西村雅史. Bert による日本語文章の難易度推定. In **IEICE Conferences Archives**. The Institute of Electronics, Information and Communication Engineers, 2022.

[5] 小木曽智信, 小町守, 松本裕治. 歴史的日本語資料を対象とした形態素解析. 自然言語処理, Vol. 20, No. 5, pp. 727–748, 2013.

[6] 小木曽智信. 旧仮名遣いの口語文を対象とした形態素解析辞書. じんもんこん 2012 論文集, Vol. 2012, No. 7, pp. 25–32, 2012.

[7] 小木曽智信, 小椋秀樹, 田中牧郎, 近藤明日子, 伝康晴. 中古和文を対象とした形態素解析辞書の開発. 研究報告人文科学とコンピュータ (CH), Vol. 2010, No. 4, pp. 1–8, 2010.

[8] 小木曽智信. 中古仮名文学作品の形態素解析. 日本語の研究, Vol. 9, No. 4, pp. 49–62, 2013.

[9] 小木曽智信, 鴻野知暁, 市村太郎. 狂言台本の形態素解析 (ブース発表, 日本語学会 2015 年度春季大会研究発表会発表要旨). 日本語の研究, Vol. 11, No. 4, p. 108, 2015.

[10] 小木曽智信, 市村太郎, 鴻野知暁. 近世口語資料の形態素解析の試み. 第 4 回コーパス日本語学ワークショップ予稿集, pp. 145–150, 2013.

[11] 森田一, 黒橋禎夫. Rnn 言語モデルを用いた日本語形態素解析の実用化. 第 78 回全国大会講演論文集, Vol. 2016, No. 1, pp. 13–14, 2016.

[12] 植田暢大, 大村和正, 児玉貴志, 清丸寛一, 村脇有吾, 河原大輔, 黒橋禎夫. Kwja: 汎用言語モデルに基づく日本語解析器. 第 253 回自然言語処理研究会, 京都, 2022.

[13] T. KUDO. Mecab : Yet another part-of-speech and morphological analyzer. **http://mecab.sourceforge.net/**, 2005.

[14] 堤智昭, 小木曽智信. 歴史的資料を対象とした複数の unidic 辞書による形態素解析支援ツール 『web 茶まめ』. じんもんこん 2015 論文集, Vol. 2015, pp. 179–184, 2015.

[15] 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, Vol. 22, No. 0, pp. 101–123, 2007.

[16] 岡照晃. Crf 素性テンプレートの見直しによるモデルサイズを軽量化した解析用 unidic: unidic-cwj-2.2. 0 と unidic-csj-2.2. 0. 言語資源活用ワークショップ発表論文集= Proceedings of Language Resources Workshop, 第 2 巻, pp. 144–153. 国立国語研究所, 2017.

[17] 岡照晃. 言語研究のための電子化辞書. コーパスと辞書, 講座 日本語コーパス, 第 7 巻, pp. 1–28. 朝倉書店, 2019.

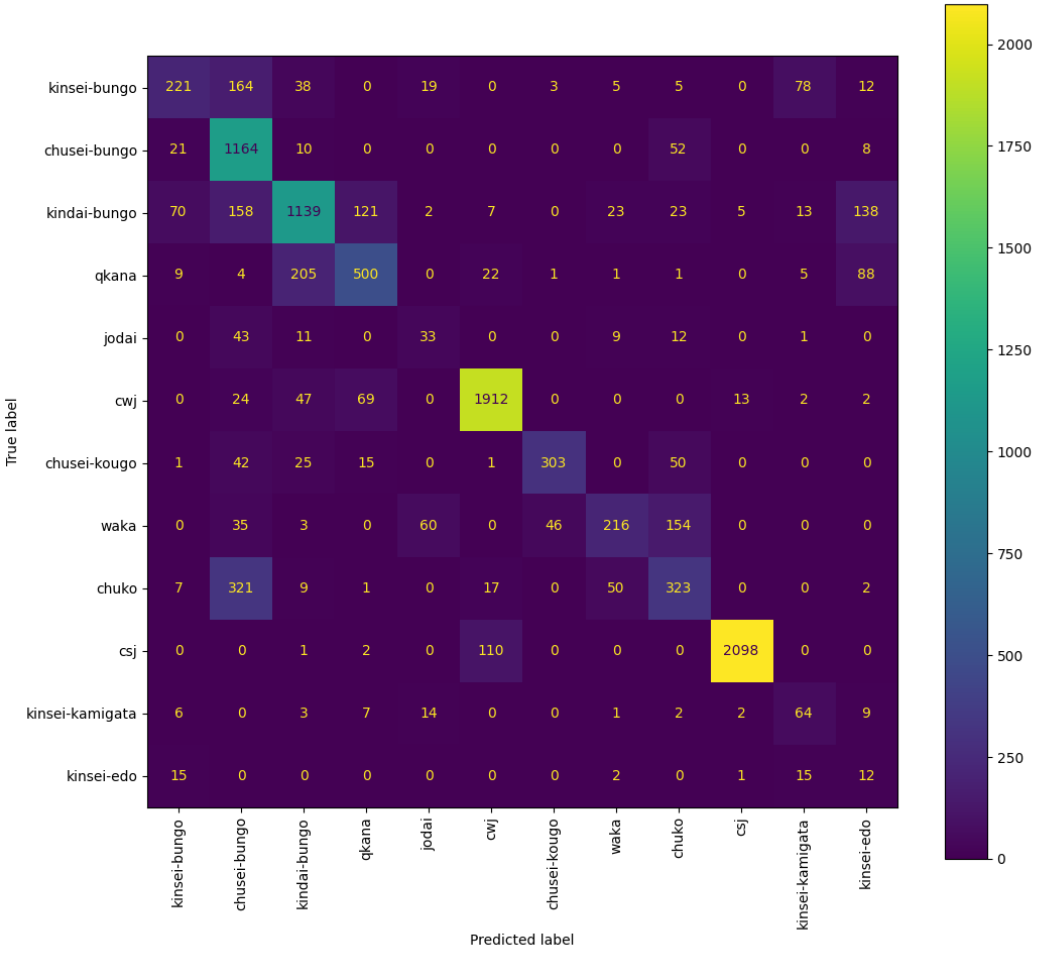

図 2　文書分類システムの出力結果の混同行列

表 3　文書分類に用いた各辞書のデータの件数

辞書名	cwj	chusei-kougo	kinsei-edo	csj	kinsei-bungo	jodai
データ数	2069	437	**46**	**2211**	546	109
辞書名	kinsei-kamigata	kindai-bungo	qkana	waka	chuko	chusei-bungo
データ数	109	1709	836	514	731	1256

表 4　文書分類システムによる辞書推定の結果 (太字：文書分類の結果正解した辞書名)

辞書名	正解	モデル 1	モデル 2	モデル 3
日本語読本 1	kindai-bungo	**kindai-bungo**	kindai-bungo	kinsei-bungo
日本語読本 2	Qkana	kindai-bungo	**Qkana**	cwj
天草版金句集	chusei-bungo	**chusei-bungo**	kinsei-bungo	**chusei-bungo**
古今集遠鏡	kinsei-kamigata, chuko	kinsei-bungo	kinsei-bungo	**chuko**