現代文 BERT を利用した日本語歴史コーパスの語義曖昧性解消

多喜 凪
東京農工大学工学部情報工学科
s182739s@st.go.tuat.ac.jp

古宮 嘉那子
東京農工大学大学院工学研究院
kkomiya@go.tuat.ac.jp

概要

本研究では，現代文 BERT による日本語歴史コーパスの語義曖昧性解消を行う．分類語彙表の語義タグが割り当てられた 8 作品 38 単語を対象とする．分析を行ったところ，訓練済み日本語現代文 BERT モデルを用いて固定のハイパーパラメータで fine-tuning した場合に，マクロ平均が 83.30%，マイクロ平均が 85.96%となり，いずれの結果もベースラインを有意に上回った．特に出現頻度が高い単語の語義曖昧性解消に対して優れた結果を示した．

1 はじめに

語義曖昧性解消（Word Sense Disambiguation, WSD）とは，自然言語処理の機械学習において，文中のある単語がどのような意味か判断する処理である．たとえば「plant」という単語の場合，「植物」や「工場」といった複数の語義が挙げられる．このような多義語の語義を，対象単語の周囲の単語の品詞や，単語同士の共起関係などを特徴として推定する．

WSD は上記の「plant」の例のように，一つの単語に対してそれぞれ別の単語が割り当てられるケースがあるような機械翻訳の分野において，非常に重要である．多義語の語義を正確に推定することで，自動翻訳や質問応答など多数のアプリケーションの精度向上に寄与することができる．

古文の WSD を行うにあたり，日本語歴史コーパス（CHJ）[1] においてはデータが少量であるため，現代文と比較して語義の識別精度が低いことが課題である．さらに，研究範囲を古文全体へ広げると奈良時代から江戸時代までが対象となり，言語に時代的な開きがある．これらのことから，様々なタスクで高い精度の古文モデルを得ることは難しいといわれている．

本研究では，BERT という汎用性の高い自然言語処理モデルによる，日本語歴史コーパスの WSD を行う．2019 年に東北大学 [2] から公開された，より高精度の訓練済み日本語現代文 BERT モデルを用いて fine-tuning する手法をとる．

2 関連研究

現代文の WSD に関する研究は，数多く行われている．なかでも近年の，単語の分散表現を利用した手法について述べる．2013 年に Tomas Mikolov ら [3, 4, 5] が自然言語処理のモデルである word2vec を公開した．これは，同じ文脈に現れる単語は類似した意味を持つという考えに基づいた手法である．word2vec の 2 層ニューラルネットワークというシンプルな構造によって，大量の文章データから単語間の意味関係をベクトルとして，現実的な計算量で得ることが可能となった．

この技術を利用した，現代日本語書き言葉コーパスの教師なし all-words の WSD として，鈴木ら [6] の研究がある．鈴木らは対象単語とその類義語から周辺単語の分散表現を作成し，ユークリッド距離の計算によって語義の推定を行っている．

古文用形態素解析辞書の開発に関しては，小木曽 [7] の研究がある．小木曽は，国立国語研究所が中心となって開発した電子化辞書 UniDic を古文へ応用し，幅広い時代の日本語に対応した通時コーパスを構築する計画を紹介している．UniDic では，見出し語に揺れの少ない短単位を採用している．さらに，テキストの種類でグループ化し，地の文と会話文を区別して辞書の作成を行うことで，古文用形態素解析辞書の解析精度を向上させることが可能であると小木曽は述べている．

高久ら [8] は，通時的な領域適応によって古文から現代文への機械翻訳を行う手法を提案した．高久らは通時適応で特定の時代にしか現れない語彙を訳出し，翻訳品質の改善が見られたと結論づけている．田邊 [9] は word2vec を用いた古文の WSD を行い，古文コーパスから作った分散表現を初期値として，現代文のコーパスで fine-tuning する手法の有効性を示している．

3 提案手法

本研究では，東北大学の日本語の訓練済み現代文 BERT モデルを利用する．BERT は，Transformer をベースにした，自然言語処理モデルである．2018 年に Google の Devlin ら [10] によって英語モデルが公開された．Devlin らは深層双方向アーキテクチャの有用性と，事前学習と fine-tuning が幅広いタスクに有効であることを示した．fine-tuning は領域適応の手法の 1 つで，初期のパラメータを再学習によって，様々なタスクに応じて微調整する．そこで，学習用の目的データが少ない場合，多量のデータを用いて事前学習を行い，その学習データを目的データへ適応させるという手段が考えられる．本研究ではこのように，現代文 BERT の古文への fine-tuning を行う．実験手法は 3 種類設定する．

(1)訓練済み現代文 BERT モデルを古文の WSD の各対象語に fine-tuning する．

(2)(1) と同様の fine-tuning を行うが，前の対象語の学習で得られたモデルを次の対象語の学習に使用する．

(3)訓練済み現代文 BERT モデルの出力を 2 つに変更し，古文の文書分類と WSD に対して，同時に fine-tuning する．

本研究で扱うモデルは BERT-base と同等のアーキテクチャをもち，2020 年 8 月時点の日本語版 Wikipedia データを使用して，事前学習を行っている．その中でも，より精度が高くなるとされる Whole Word Masking を導入したバージョンを用いる．Whole Word Masking とは，事前学習時に単語単位で MASK を行い，MASK する単語に対応するサブワードも MASK する方式である．

4 データ

4.1 コーパス

日本語歴史コーパスは，国立国語研究所によって日本語史研究の基礎資料としての開発のために構築されているコーパスである．全テキストに読み・品詞などの形態論情報が付与されている．電子資料として，総索引ができることに加え，より高度な検索や集計を行うことが可能となる．

本研究では，竹取物語，土佐日記，今昔物語集，方丈記，宇治拾遺物語，十訓抄，徒然草，虎明本狂言の 8 作品の日本語歴史コーパスを用いる．作品と単語数について表 1 に示す．平安時代と鎌倉時代の作品が中心となる．コーパスの 10 項目のうち，本研究では出現書字形（orthToken），語彙素（lemma），分類語彙表番号（Word List by Semantic Principles,wlsp）の 3 項目を主に利用する．コーパス名（corpusName），文境界（boundary）の 2 項目は補助的に利用する．

表 1 日本語歴史コーパスの作品と単語数

作品名	単語数	時代	ジャンル
竹取物語	12758	平安	物語
土佐日記	8209	平安	日記
今昔物語集	175602	平安	説話
方丈記	5403	鎌倉	随筆
宇治拾遺物語	120706	鎌倉	説話
十訓抄	90178	鎌倉	説話
徒然草	40835	鎌倉	随筆
虎明本狂言	5449	室町	能狂言

分類語彙表 [11] とは，語を意味によって分類および整理したシソーラス（類義語集）である．分類番号を用いることで，上位の概念や同じ概念の単語を容易に導くことができる．分類語彙表のレコード総数は 101,070 件で，一つのレコードの構成は「レコード ID 番号／見出し番号／レコード種別／類／部門／中項目／分類項目／分類番号／段落番号／小段落番号／語番号／見出し／見出し本体／読み／逆読み」となっている．分類番号は「類／部門／中項目／分類項目」で構成される．

例えば「手」という語は分類語彙表の複数箇所に出現し，1.1401, 1.3081 などの分類番号がある．1.1401 の語は「手」のほかに「労働力」「男手」などがある．1.3081 の語に，「方法」「すべ」などがある．分類番号は，単語の語義として扱うことができる．

4.2 実験データ

日本語歴史コーパス中において分類番号が付与され，かつ 500 回以上出現があった 38 単語を対象語とする（表 2）．対象語はいずれも多義語である．対象語を含む 1 文を，WSD を行う語に関する 1 つの入力としてデータを作成する．虎明本狂言以外は句点（。）を文末として，文を区切る．なお，台本形式の虎明本狂言のみ，文中に句点が出現しない．そのため，虎明本狂言のみ「および『』，boundary で B（区切り）に該当した読点（、）を文末とみなす．他作品と条件をそろえるため，文末とみなした記号は

表 2 8 作品の対象語 38 単語

為る	言う	有る	事	人	見る	無い
是	物	思う	行く	時	取る	其
出でる	然る	此れ	者	様	成る	所
知る	申す	程	後	家	持つ	返る
心	居る	来る	間	立つ	国	女
参る	日	下				

表 3 epoch と lr の設定

実験手法	epoch	lr
(1)	10	0.0001
(2)	10	0.0001
(3)	10	0.0001
(1) 調整	10,20,30	0.00001,0.0001,0.001

句点に置き換えた後，分かち書きや Encode を行う．

日本語歴史コーパスは古語特有の語義を含むことから，未知語の割合が現代語より高くなると考えられる．すべての単語を Encode した中の未知語の割合は，8 作品で 17.06%となっている．

次に，対象語の抽出方法について述べる．日本語歴史コーパス中の lemma の項目で完全一致する場合のみ，BERT への入力文に採用する．lemma が基準であるため，動詞では基本形が一致していれば活用形が異なっていても該当する．1 文に 2 回以上対象語の完全一致が見られた場合，それぞれを対象語の位置が違う別の入力文として扱う．

5 実験

5.1 実験設定

新納の書籍 [12, 13] を参考に実験プログラムを作成する．対象語を含む 1 文を，語義を判定したい語に関する 1 つの入力として扱う．次に，対象語の語義数を集計し，語義ごとに 0 から整数の番号をあてる．入力文はランダムに並び替え，時代も作品もシャッフルされた状態にする．対象の 1 文について UniDic 基準で分かち書きを行い，BERT によって Encode する．本研究で使用した BERT 自体の機能を用いて分かち書きを行う場合は NEologd が採用されることになる．しかし，データとなる日本語歴史コーパスにあらかじめ，各時代に対応した UniDic 基準の形態論情報が付与されているため，本研究の分かち書きではそちらを採用する．文中の対象語の位置に先頭から数えた整数の番号をあてる．

最適化関数は SGD，損失関数はクロスエントロピー誤差を用いる．ハイパーパラメータについては表 3 で示す．速度と正答率の観点から事前実験を行い，epoch（学習回数）を 10，lr（学習率）を 0.0001 とする．データセットは 4：1 で訓練データ（train）とテストデータ（test）に分ける．さらに手法 (1) について，ハイパーパラメータをグリッドサーチで調整する．その際のデータセットの分け方は，3：1：1（訓練データ：開発データ：テストデータ）とする．3 通りの epoch と lr を掛け合わせ，9 通りのモデルを作成する．その中から，開発データでの正答率が高いモデルを対象語ごとに採用する．

5.2 評価手法

WSD は，1 文の入力に対し正解のラベルを持っている．テストデータにおいて，入力文の正解ラベルと BERT が予測したラベルが一致した場合を正解とする．正答数を問題数で割ったマクロ平均とマイクロ平均を正答率とする．ここでのマクロ平均は，各単語の正答率の平均である．マイクロ平均は，全単語の正答数を問題数で割った正答率である．

なお，例えば 4 種類の語義があるとして，それらがすべて訓練データにもテストデータにも含まれているという保障はない．ランダムな文の並べ替えによって，どちらかのデータにしか出現しない語彙となっている可能性がある．2 出力時は WSD と文書分類の同時学習となるが，比較のため文書分類のみの BERT の fine-tuning も行う．表 3 で示した手法とベースラインのマイクロ平均について，カイ二乗検定を行い，手法同士の有意差を調べる．

最頻出語義（Most Frequent Sense, MFS）をベースラインに設定する．最頻出語義の確率は，その対象語に関して 1 番多く出現する語義を予測ラベルとした時の，正答率を指す．38 単語の平均語義数は 8.21 個となっている．wlsp が 1 つの単語に 2 つ設定されている場合や空欄である場合も，それ自体を 1 つの語義とみなし，語義数にカウントする．2 つ設定されているのは，2 つの語義のどちらともとれる単語，空欄は語義が不明な単語であると考えられる．また，空欄の語義が MFS や 2 番目に多く出現する語義となっている対象語がある．このように曖昧な状態も数多く判別する必要がある．

6 実験結果

実験結果を表 4〜7 に示す．いずれの提案手法も，ベースラインを有意に上回った（表 4）．実験結果

表 4　WSD の実験結果

	マクロ平均	マイクロ平均
MFS	62.04	57.99
(1) 通常	**83.30**	**85.96**
(2) モデル更新	82.37	85.49
(3) 2 出力	81.82	85.32
(1) 調整	81.75	84.93

表 5　各対象語の WSD の実験結果

対象語	(1)	MFS	対象語	(1)	MFS
為る	59.27	24.12	言う	**94.37**	50.21
有る	**95.34**	57.09	事	**95.11**	58.74
人	89.15	49.67	見る	**90.00**	87.99
無い	**97.83**	**97.88**	是	**99.83**	55.70
物	76.64	55.89	思う	**95.23**	54.71
行く	**94.22**	**94.22**	時	86.15	76.30
取る	60.62	46.63	其	**99.81**	58.04
出でる	81.25	79.55	然る	67.92	31.17
此れ	89.72	51.10	者	88.89	88.79
様	80.16	56.53	成る	61.83	48.39
所	79.47	77.03	知る	88.89	89.06
申す	86.84	65.09	程	82.23	50.04
後	**91.72**	**91.93**	家	83.22	79.78
持つ	82.22	46.52	返る	77.69	76.15
心	84.10	54.00	居る	62.94	60.36
来る	**93.90**	70.78	間	74.59	69.95
立つ	80.87	71.33	国	84.21	62.74
女	**96.86**	44.37	参る	68.97	63.39
日	71.24	32.33	下	71.94	29.87

から，現代文の BERT が古文の WSD に有用であることが分かる．最も正答率が高かったのは，数字を太字で示した箇所で，手法 (1) となった．(1) をハイパーパラメータ調整した場合には，マクロ平均で 81.75%，マイクロ平均で 84.93%となった．

表 4，6 より，38 単語総合の WSD に関してはすべての手法において，マイクロ平均がマクロ平均を上回った．表 5 は，対象語ごとの手法 (1) の WSD の実験結果である．正答率が 90%以上となった箇所を太字で示す．対象語の配置は，為る，言う，有る……の順に，分類番号を持つ単語としてのコーパス中の出現回数が多く，降順になっている．表 7 では，最も正答率が高い箇所と最もテストデータが多い箇所を太字で示す．

表 6　2 出力の WSD の実験結果

	マクロ平均	マイクロ平均
2 出力 WSD	81.82	85.32
WSD のみ	83.30	85.96
2 出力 文書分類	88.81	89.04
文書分類のみ	89.66	89.62

表 7　各作品の WSD の実験結果

	(1)	テスト データ数
竹取物語	75.50	347
土佐日記	83.65	208
今昔物語集	**88.39**	**5091**
方丈記	76.03	121
宇治拾遺物語	85.52	3280
十訓抄	83.09	1969
徒然草	80.44	992
虎明本狂言	70.79	89

7　考察

マクロ平均はデータが少ない事例の影響を強く受ける．マクロ平均がマイクロ平均より常に低いため，データが少ない場合に正答率が低い傾向にあると推測できる．表 6 からも，出現回数が多い表上部の単語に 90%以上の正答率が多くみられるため，BERT の WSD では学習データが増えるほど，正答率が高くなると考えられる．

表 4，6 より，手法 (2) の単語を超えた重み共有や手法 (3) の文書分類タスクとの重み共有は，本実験では全体の正答率を上昇させる条件とならなかったことが分かる．本実験の対象語は名詞や動詞が区別なく交ざっていた．また表 7 でも，時代による特徴は明確にはみられない．また，今回のデータは鎌倉時代のデータが多く，時代の影響は比較的少なかった可能性がある．よって，対象語同士や作品同士を関連づけて学習できていないと考えられる．

8　おわりに

本研究では，現代文 BERT による日本語歴史コーパスの WSD を行った．実験により，現代文の BERT は古文の WSD に有効であることを示した．また，単語を超えた重み共有と文書分類タスクとの重み共有は有効でないことが分かった．本実験の対象語は名詞や動詞が交ざっていたこと，作品や時代による影響が比較的少なかったことが原因と考えられる．

謝辞

本研究は JSPS 科研費 17H00917，17KK0002，18K11421 の助成を受けたものです．また，国立国語共同研究プロジェクト 「通時コーパスの構築と日本語史研究の新展開」および「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」の成果です．データセットの作成にあたり，2022 年 1 月 8 日時点で最新の語義タグ付き日本語歴史コーパスを提供してくださった，浅原正幸教授（国立国語研究所）に感謝いたします．

参考文献

[1] 国立国語研究所. 『日本語歴史コーパス』2022 年 1 月 8 日確認. https://ccd.ninjal.ac.jp/chj/.

[2] 東北大学乾研究室. bert-japanese. 2019. https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/.

[3] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. **Proceedings of NAACL-HLT**, pp. 746–751, 2013.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. **Proceedings of ICLR**, 2013.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. **Proceedings of NIPS**, 2013.

[6] 鈴木類, 古宮嘉那子, 浅原正幸, 佐々木稔, 新納浩幸. 概念辞書の類義語と分散表現を利用した教師なし all-words wsd. 自然言語処理, Vol. 26, No. 2, pp. 361–379, 2019.

[7] 小木曽智信. 通時コーパスの構築に向けた古文用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. 6, No. 2011-CH-92, pp. 1–4, 2011.

[8] 高久雅史, 平澤寅庄, 小町守, 古宮嘉那子. 通時的な領域適応を行った単語分散表現を利用した古文から現代文へのニューラル機械翻訳. 言語処理学会第 26 回年次大会, 2020.

[9] 田邊絢. 現代文と古文の情報を利用した日本語歴史コーパスの語義曖昧性解消の領域適応. 茨城大学大学院理工学研究科 2019 年度修士学位論文, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **Proceedings of NAACL-HLT**, 2019.

[11] 国立国語研究所. 『分類語彙表』2022 年 1 月 8 日確認. https://ccd.ninjal.ac.jp/goihyo.html.

[12] 新納浩幸. Pytorch 自然言語処理プログラミング. インプレス, 2021.

[13] 新納浩幸. Pytorch による物体検出. オーム社, 2020.