日本語歴史コーパスの All-words WSD

浅田宗磨
東京農工大学工学部
s197006x@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，日本語歴史コーパスの語義曖昧性解消を行った．日本語歴史コーパスの語義曖昧性解消の研究には，頻出語のみを対象とする lexical sample task を一つの単語ごとにモデルを作成する手法で解いた [1] があるが，本研究では系列ラベリングの手法で全単語について一つのモデルを学習させ，all-words の語義曖昧性解消を行った．また，単語に割り当てられた分類語彙表による語義の概念の粒度を変更して 2 通りの実験を行った．実験の結果，どちらの実験でも提案する手法が最頻出語義のベースラインを有意に上回った．

1 はじめに

語義曖昧性解消（Word Sense Disambiguation，WSD）とは，文中の多義語がどのような語義か判断する処理である．多義語とは，複数の語義を有している単語のことで，たとえば「夢」という単語には「睡眠時にみられる現実に似た一連の観念や心像」という語義や「実現させたいと思っているがまだ空想の域を出ない事柄」という語義がある．

多義語の語義は，文脈によって判断される．機械学習においては，文脈情報として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語義を推定する．WSD は，文中の単語の語義を示すため，初学者に向けた文章読解や単語学習に役立てられる．

WSD には，コーパス中に頻出する単語だけを対象にする lexical sample task と，コーパス中の全単語を対象にする all-words WSD がある．一般に，lexical sample task では単語ごとに分類モデルを学習する方法をとるのに対して，all-words WSD では系列ラベリングの手法で全単語についてひとつのモデルを学習させることが多い．古文の WSD を行うにあたり，日本語歴史コーパス（CHJ）[2] においては語義タグ付きデータが少量であったため，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った．しかし，2022 年の前半に CHJ の語義タグのタグ付け作業が終了したため，現在では多くのデータが手に入るようになった．

日本語歴史コーパスを対象とした WSD の先行研究としては，lexical sample task として解いた [1] があるが，本研究では同じデータに対して all-words WSD を行う．このことにより，コーパス中にあらわれる頻度の低い多義語についての語義を推定することが可能になる．また，本研究では，語義として利用されている，分類語彙表の概念の粒度を変更して実験を 2 通り行った．

2 関連研究

all-words WSD の先行研究には，鈴木ら [3] の研究がある．この研究では対象単語の周辺単語の分散表現を作成し，類義語の周辺単語の分散表現とのユークリッド距離を計算することで対象単語の語義を予測し，その手法の有効性を示した．ただし，対象文書は古文ではなく，現代日本語書き言葉均衡コーパス [4] である．また，WSD に古文と現代日本語文の通時的な領域適応を行った論文として，Komiya ら [5] の研究がある．この研究では，古文の WSD のための素性として様々な種類を比較し，古文によって作成された分散表現に現代文による fine-tuning を行った素性が古文の WSD に有効であると示した．

BERT[6] を用いた英語の all-words WSD に，Jiaju ら [7] の研究がある．彼らは，BERT を encoder として利用し，得られた素性が all-words WSD に有効であると示した．日本語の分散表現の研究に，新納ら [8] の研究がある．この研究では，国語研日本語ウェブコーパス [9] と word2vec[10] を用いて作成した分散表現が日本語 WSD に有効であると示した．all-words WSD の手法で同形異音語の読み推定を行った研究に，小林ら [11] の研究がある．この研究では，単語の「読み」を「語義」とみなして，BERT の fine-tuning を用いて読み推定を行った，ま

た，日本語歴史コーパスの頻出単語の WSD として，Komiya ら [1] の研究がある．彼らは訓練事例の不足を，現代文を大量に利用して学習した BERT を利用することで補い，これを新たにタグ付けされた古文の WSD 用のデータで fine-tuning することで，高い性能を得られることを示した．先行研究は lexical sample task であるが，本研究では同じデータを用いて all-words WSD を行った．

3 提案手法

本研究では，先行研究 [1] にならい，対象コーパスは古文であるが，現代文の BERT を事前学習モデルとして利用し，これを古文の語義タグつきデータで fine-tuning することで WSD を行った．また，本研究では，all-words WSD を，入力文中の WSD の対象単語すべてに語義タグを付与する，系列ラベリングのタスクとしてとらえ，実装を行った．ただし，固有表現抽出などの一般的な系列ラベリングタスクと違って，all-words WSD では単語ごとに候補となる語義ラベルの集合が異なる．例えば，「犬」という単語に「犬」以外の単語の語義ラベルを付与するのは望ましくない．そのため，WSD の対象単語ごとに候補となる語義ラベル集合を参照し，その候補の語義ラベル集合の中で最も softmax 関数の出力が高いものを正解とみなして学習及び推論を行った．

4 実験

本研究では，東北大学が公開した日本語の訓練済み現代文 BERT モデル[1)]を利用する．本研究で用いる BERT モデルは，BERT-base と同等のアーキテクチャで，2020 年 8 月 31 時点の日本語 Wikipedia から作成された約 3,000 万文のコーパスを用いた事前学習が行われたものである．本研究ではこの現代文 BERT を fine-tuning する．

実験は，分類語彙表番号が最も詳細な 5 桁の番号と，上 3 桁のみ残し他は切り捨てたものの 2 種類を行う．表 1 から，分類語彙表番号の桁を 5 桁から 3 桁にすることで，曖昧性が減少し，対象語義種類数が大幅に減少していることが分かる．また，WSD の対象単語出現数と，対象単語種類数もともに減少している．結果として，対象単語平均語義数は 2.91 と 2.73 であり，対象語義種類数と比較すると，小さな差となっている．

1) https://huggingface.co/cl-tohoku/bert-base-japanese-v2

表 1 日本語歴史コーパスの統計情報

	5 桁	3 桁
総単語出現数	647,751	647,751
対象単語出現数	329,109	324,952
対象単語種類数	3,878	3,672
対象語義種類数	1,747	304
対象単語平均語義数	2.91	2.73

4.1 実験設定

システムへの入力は文とした．この際，日本語歴史コーパス中の句点（。）を文末として区切られる 1 文を入力とする．虎明本狂言は台本形式のため，句点が文中に出現しない．そのため，鍵括弧（「，『），文境界で B（区切り）が割り当てられた読点（、）を，句点に置換した上で文末として 1 文を区切る．入力文中の各単語の素性として，日本語歴史コーパスの出現書字形，語彙素，分類語彙表番号を用いる．入力文は時代や作品によらずランダムな順序で学習を行った．日本語歴史コーパスの単語区切りは Unidic[12] の区切りとなっているため，BERT には Unidic の区切りで encode する東北大学の bert-base-japanese-v2 を利用した．しかし，BERT の語彙は限られているため，BERT encoder により文を分割すると，日本語歴史コーパスの単語が複数に分割されることがある．例えば，「痛み入る」は日本語歴史コーパス中では 1 つの単語として扱われているが BERT encoder によって「痛み／入る」のように分割される．こうした場合には，最初のトークンを WSD の対象単語として利用した．

また，学習時には最適化関数に Adam を，損失関数にクロスエントロピー誤差を用いた．ハイパーパラメータは表 2 に示す値のグリッドサーチを行い，開発データにおいて最も正解率の高い値を採用した．訓練データ: 検証データ：テストデータは 3:1:1 として五分割交差検定を行った．

表 2 ハイパーパラメータの調整

分類語彙表番号	epoch	lr
5 桁	5,10,15	3e-5,1e-5,3e-6
3 桁	10,15	3e-5,1e-5,3e-6

4.2 評価手法

コーパス中の全多義語について，以下の式を用いて正解率を求めた．

$$\text{正解率} = \frac{\text{正解単語数}}{\text{全多義語数}} \quad (1)$$

なお，ある単語がコーパス中で複数の語義を有しているとしても，それらがすべて訓練データにもテストデータにも登場するとは限らない．日本語歴史コーパス中で 2 回しか登場しない多義語はコーパス中の多義語の 1 割程度を占める．

本実験では，最頻出語義（Most Frequent Sense, MFS）をベースラインに設定した．MFS は以下の式で求められる．

$$\text{対象単語の MFS} = \frac{\text{最頻出語義出現数}}{\text{全出現回数}} \quad (2)$$

5 コーパス

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，太陽，尋常小学読本（1904,1910）の計 10 作品の日本語歴史コーパスを用いる．コーパスの統計情報は表 1 の通りである．

分類語彙表 [13] とは，語を意味によって分類・整理したシソーラス（類義語集）である．分類語彙表は似た概念を持つ単語や，概念の包含関係などの情報が扱いやすい形態で記録されている．分類語彙表中の一つのレコードは「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」の項目からなる．分類番号はこのうち，「類／部門／中項目／分類項目」と対応する．たとえば「言う」という単語に 2.3102 という番号が振られている場合，各項目との対応は「2.／3／1／02」となる．分類番号は位の大きさが小さくなるにつれ分類がより詳細になっている．

6 実験結果

実験結果を表 3 に示す．分類語彙表番号が 5 桁，3 桁の双方において正解率がベースラインを上回った．この差は，自由度を 0.05 としてカイ二乗検定を行うと有意であることがわかった．実験結果から，現代文 BERT は古文の all-words WSD にも有効であることがわかる．

表 3 WSD の実験結果

分類語彙表番号	実験手法	正解率 (%)
5 桁	MFS	81.61
	提案手法	84.52
3 桁	MFS	84.10
	提案手法	87.11

コーパス内の頻出語のみを対象語とした lexical sample task の WSD においても，学習データ数が少ない際には MFS を超すのは難しいとされる（[1], [5]）．また，今回の研究と同じコーパス中に 1,000 回以上出現する 33 単語を対象語とした lexical sample task の 5 桁の分類語彙番号を使用した際の正解率 [1] はマクロ平均で 84.68%，MFS が 78.29%であった．比較すると，本研究で利用した全単語のデータの方が 3.32 ポイント MFS が高いため，コーパス中に出てくる頻度の小さい単語のほうが MFS が低いことが分かる．しかし，一般に機械学習においては学習データが少ない場合，正解率が低い傾向にある．そのため，MFS を 2.91 ポイント及び 3.01 ポイント上回った今回の結果は十分高い結果だと考えている．

表 4 出現頻度の小さい場合の正解率 (%)

出現回数	5 桁	MFS	3 桁	MFS
2	51.91	50.00	53.43	50.00
3	51.13	63.35	55.43	63.98
4	51.08	62.92	58.48	63.66
5	53.56	67.06	59.81	67.80
6	55.98	67.21	61.69	69.11
7	59.62	69.43	62.40	70.74
8	59.48	71.22	61.73	72.01
9	61.66	70.87	66.24	70.11

表 5 誤答が頻出した単語

順位	5 桁	3 桁
1	為る	為る
2	成る	成る
3	然る	然る
4	物	物
5	取る	取る
6	人	共
7	様	様
8	共	ばかり
9	中	皆
10	時	又

7 考察

コーパス中での出現頻度の小さい単語での正解率を表 4 に示す．分類語彙表番号が 5 桁，3 桁の両方の場合において，正解率は出現回数が 2 回であるものを除いて MFS を大きく下回った．これは，6 節で述べた，学習データ数が少ない場合の極端な例である．この結果は，対象単語が訓練データに出現せず，学習による語義推定の手がかりが得られない確率が高いためだと考えられる．

また，システムがよく誤答した 10 単語を上から頻出順に表 5 に示す．システムが誤答した単語で多かったのは，「為る」「成る」「然る」といったコーパス中に頻出で，語義数が 10 を超えるものであった．一方で，また，語義の概念の粒度によってよく誤答する単語に違いがみられた．たとえば「人」という単語は，分類語彙表番号が 5 桁の場合では 7 種類の語義に分類される中で 484 回誤答していた．3 桁の場合では 6 種類の語義に分類される中で 123 回誤答していた．この結果から，「人」という単語は微妙で学習しづらい細分化された概念を有していると考えられる．

8 おわりに

本研究では，日本語歴史コーパスの all-words WSD を行った．語義として利用した分類語彙表番号を 5 桁すべて利用する場合と，上位 3 桁までを利用した場合の 2 種類の実験を行った．実験により，両方の場合において，現代文 BERT は古文の all-words WSD に有効であることがわかった．コーパス中の出現頻度が小さい単語については，MFS と比較して高い精度を得られなかった．また，語義の概念の粒度が異なると，一部の単語において語義数に違いがあまり見られないのにもかかわらず誤答が多くなることがわかった．

謝辞

本研究は JSPS 科研費 17KK0002，18K11421，22K12145 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「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」の成果です．

参考文献

[1] Komiya Kanako, Oki Nagi, and Asahara Masayuki. Word sense disambiguation of corpus of historical japanese using japanese bert trained with contemporary texts. In **PACLIC 2022**, 2022.

[2] 国立国語研究所. 『日本語歴史コーパス』バージョン 2022.3, 2022. https://clrd.ninjal.ac.jp/chj/.

[3] 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.

[4] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language Resource and Evaluation**, Vol. 33, No. 7, pp. 345–371, 2014.

[5] Komiya Kanako, Tanabe Aya, and Shinnou Hiroyuki. Diachronic domain adaptation of word sense disambiguation for corpus of historical japanese using word embeddings. **NINJAL Research Papers**, Vol. 23, pp. 29–57, jul 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Jiaju Du, Fanchao Qi, and Maosong Sun. Using bert for word sense disambiguation. **arXiv preprint arXiv:1909.08358**, 2019.

[8] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木稔. nwjc2vec: 国語研日本語ウェブコーパスから構築した単語の分散表現データ. 自然言語処理, Vol. 24, No. 5, pp. 705–720, 2017.

[9] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. Archiving and analysing techniques of the ultra-large-scale web-based corpus project of ninjal, japan. **Alexandria**, Vol. 26, No. 1–2, pp. 129–148, 2014.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, **Advances in Neural Information Pro-**

cessing Systems 26, pp. 3111–3119. Curran Associates, Inc., 2013.

[11] 小林汰一郎, 古宮嘉那子, 新納浩幸. 疑似訓練データを用いた bert による同形異音語の読み推定. 第 253 回自然言語処理研究発表会, 2022.

[12] 康晴伝, 智信小木曽, 秀樹小椋, 篤山田, 信明峯松, 清貴内元, 花絵小磯. 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, pp. 101–123, oct 2007.

[13] 国立国語研究所. 『分類語彙表増補改訂版データベース』(ver.1.0), 2004. https://clrd.ninjal.ac.jp/goihyo.html.