『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号悉皆付与

浅田 宗磨
東京農工大学
s231157v@st.go.tuat.ac.jp

古宮 嘉那子
東京農工大学
kkomiya@go.tuat.ac.jp

浅原 正幸
国立国語研究所
masayu-a@ninjal.ac.jp

概要

本研究では、コーパスを語義情報で検索できるようにすることを目標とし、『現代日本語書き言葉均衡コーパス』(BCCWJ) に対して『分類語彙表』の語義情報を悉皆付与した。BCCWJ-WLSP を訓練データとした all-words WSD モデルを構築し、BCCWJ 1 億語規模の分類語彙表番号付きデータを構築した。訓練に用いなかった 10 レジスタについて、500 語に対して人手による性能評価を行い、データの有用性を検証した。さらに、構築した語義付き語彙表について解説する。

1 はじめに

『現代日本語書き言葉均衡コーパス』(BCCWJ)[1] は検索系「中納言」により、文字列、正規表現や形態論情報に基づく検索を行うことができる。品詞連接や活用形に基づく表現の問い合わせが可能であるが、語義情報に基づく問い合わせは行えない。本研究では、多様なレジスタからなる 1 億語規模 の BCCWJ を語義情報で検索できるようにすることを目標とし、all-words WSD 技術を用いて BCCWJ 全体に語義タグとして『分類語彙表』[2] の分類番号を付与した。本稿では、1 億語規模の語義情報付与作業について、利用したデータと解析手法を解説する。解析モデルは学習データを訓練データ・検証データ・テストデータに基づいて最良のものを用いたが、訓練データにないレジスタのデータについてさらに 500 語ずつ人手で語義情報を確認し、解析結果の性能を検証した。言語研究に資するために同データに基づき語義タグつき頻度表を構成した。語義タグに基づくレジスタ差を検証するとともに、新聞記事レジスタについては記事分類についての語彙分布の差についても検証した。

2 作業の内容

2.1 利用したデータ

『分類語彙表』[2] は区切り記号を含めて 101,170 エントリからなる語義情報付き辞書である。語義情報は分類番号と呼ばれる 5 ケタの数字で表現される。例えば、分類番号 1.1920 は、「体-関係-量-程度・限度」を表す。 この分類番号を語義タグとして用いる。分類番号のうち、ピリオドの左の数字は類と呼び、「1. 体」・「2. 用」・「3. 相」・「4. 他」の 4 つに分類される。またピリオドの右 1 ケタの数字は部門と呼び「.1 関係」・「.2 主体」・「.3 活動」・「.4 生産物」・「.5 自然」の 5 つに分類される。

『現代日本語書き言葉均衡コーパス』(BCCWJ)[1] は短単位で 1.2 億語規模からなる日本語コーパスである。コーパスは生産実態サンプルに相当する書籍 (PB)・雑誌 (PM)・新聞 (PN)、流通実態サンプルに相当する書籍 (LB)、特定目的サンプルに相当する白書 (OW)・教科書 (OT)・広報紙 (OP)・ベストセラー (OB)・Yahoo! 知恵袋 (OC)・Yahoo! ブログ (OY)・韻文 (OV)・法律 (OL)・国会会議録 (OM) の 13 レジスタからなる。このうち PB, PM, PN, OW, OC, OY の一部約 120 万語はコアと呼ばれ、人手による形態論情報が付与されている。

BCCWJ-WLSP[3, 4] はコアデータのうち PB, PM, PN の約半分のサンプルについて、人手で『分類語彙表』の語義情報を付与したものである。BCCWJ-WLSP の統計情報を表 1 に示す。今回、BCCWJ-WLSP を学習用データとして用いるにあたり、各入力文の先頭にその文の出典情報を示す title_id を付与する。title_id はコーパス中の「pSampleID」からレジスタ情報を含む先頭 4 文字を用いて、それをひとつのラベルとした。具体的には

「PN3g_00001」となっているところを「PN3g」のようにした。

表 1 BCCWJ-WLSP の統計情報

総出現単語数	347,094
対象単語出現数	126,286
対象単語種類数	2,988
語義種類数	918
出現単語平均語義数	2.55
title_id 数	75

UniDic2WLSP[5] は BCCWJ に付与された形態論情報 (UniDic 品詞体系) の語彙素番号に対して、可能な『分類語彙表』の語義情報を枚挙したデータである。本研究で、対象単語の候補となるラベルの展開に用いた。

2.2 解析手法

本研究では系列ラベリング手法で all-words WSD を行った。ただし、固有表現抽出や品詞タグ付けなどの一般的な系列ラベリングタスクとは異なり、all-words WSD では対象とする単語によって候補となるラベルの種類や数が異なる。そのため、単語ごとの候補となる語義をまとめた集合を予め作成し、システムが参照できるようにした。たとえば、「召す」という単語に「召す」が取りうる語義以外の語義を推論に入れる必要はない。そこで、「召す」が取りうる語義を予め収集して語義候補を立てることで、効率的な推論が可能になる。また、文の出典レジスタを示す ID を入力トークンの先頭に加えた。各単語の語義候補の中で最も高い出力値のものを正解とみなして学習およびタグ付けを行った。

BCCWJ を対象にした all-words WSD の先行研究に Suzuki et al.[6] の研究がある。この研究では、文中での単語の語義は文脈によって決定するという考えの下、対象単語の周辺単語の分散表現と、類義語の周辺単語の分散表現を比較し、ユークリッド距離を計算することで対象単語の語義を決定し、その有効性を示した。本研究では大規模言語モデルである BERT[7] を用いて all-words WSD を行った。BERT を用いた all-words WSD の研究に Asada et al.[8] の研究がある。この研究は『日本語歴史コーパス』[9] を対象とし、BERT が all-words WSD に有効であることを示した。更に、入力文の先頭に出典作品を示す id を付与することでシステムの正解率が向上することを示した。

本研究では先行研究 [8] に倣い、システムへの入力は文単位で行った。コーパス中で句点（「。」）あるいは文境界で区切られる一文を入力する。学習時には 510 トークンを超える文については最初の 510 トークンのみを入力とする。語義タグ付与時には超過した分のトークンは別の文として入力する。入力文をつくる各トークンは出現書字形をトークナイズしたものである。トークンを BERT encoder により id 化するとき、BCCWJ では一単語として扱われている語が複数の単語に分割される場合がある。たとえば「盛り上がる」という語が BERT encoder により「盛り／上がる」のように分割される。この場合、先頭のトークンのみを入力とする。

本研究では BERT の日本語モデル[1)] を WSD タスクに向けて fine-tuning した。BCCWJ-WLSP を fine-tuning 用の学習コーパスとして用いた。学習コーパスのうちから句点（「。」）あるいは文境界で区切られるトークン列を入力文とし、文のコーパスでの出現順序を無作為にシャッフルして五分割したものをデータセットとして扱った。そのうちの 3/5 を訓練データ、1/5 を検証データ、1/5 をテストデータとした。学習時は、入力データ列とそれに対応した語義候補集合の列および正解語義列を用いて、入力データ列中の各トークンの予測語義と正解語義を比較して学習を進めた。なお、title_id トークンは「語義無し」の単義語として扱った。学習率、最大エポック数を変えてモデルを作成し、それぞれの学習結果の中から最良の正解率 (正解率 A) をテストデータで示したモデルを実際の語義タグ付与に用いた。正解率 A は以下の式で計算される。タグ付け自体は単義語にも行ったが、正解率 A を算出するときには単義語は含めなかった。

$$\text{正解率 A} = \frac{\text{対象単語正解数 (多義語のみ)}}{\text{対象単語出現数 (多義語のみ)}} \quad (1)$$

学習時には損失関数にクロスエントロピー誤差を、最適化関数に Adam を用いた。最良モデルのテストデータでの正解率は、88.05%であった。語義タグ付与時には、作成したモデルを読み込み、BCCWJ に出現する全単語に語義タグを付与した。

3 評価

BCCWJ に全体に付与した語義タグの性能を評価するために、学習に用いなかった 10 レジスタにつ

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

いて語義タグの正誤を検証した。助詞・助動詞などの語義タグが付与されていない単語も含めて 500 語ずつ人手により語義タグの確認作業を行った。表 2 に以下に示す 2 種類の正解率を示す。

$$\text{正解率 B} = \frac{\text{対象単語正解数 (単義語も含む)}}{\text{対象単語出現数 (単義語も含む)}} \quad (2)$$

$$\text{正解率 C} = \frac{\text{全単語正解数 (タグなしも含む)}}{\text{全単語}} \quad (3)$$

正解率 B は語義付与対象である主に自立語のみを分母とした評価である。正解率 C は語義付与対象外である付属語・句読点も含めた全単語 500 語を分母とし、語義付与対象外の単語に対して語義を付与しないことについても正解とした評価である。

表 2　人手による評価結果 (各レジスタ 500 語)

レジスタ	正解	誤り	正解率 B	正解率 C
LB (書籍)	238	31	88.5%	93.8%
OB (書籍)	196	44	81.6%	91.2%
OC (知恵袋)	166	58	74.1%	88.4%
OL (法律)	264	44	85.7%	91.2%
OM (議事録)	216	54	80.0%	89.2%
OP (広報紙)	241	51	82.5%	89.8%
OT (教科書)	177	59	75.0%	88.2%
OV (韻文)	204	95	68.2%	81.0%
OW (白書)	261	36	87.9%	92.8%
OY (ブログ)	197	46	81.1%	90.8%

なお、OV (韻文) については、14 件 (2.8%) が形態素解析誤りに基づく語義タグ誤りであった。

データを確認したところ、異なり 201,797 語中 172,819 語が 語義タグ 1.1960 「体-関係-量-数記号 (一二三)」が付与されていることが分かった。これは学習データにおける未知語に対して、多様な単語に連接する数詞相当の 1.1960 が付与されたからだと考える。延べ語数にすると 124,100,965 語中の 10,250,986 語に相当するが、このうちの 3,789,581 語が品詞「名詞-数詞」に相当し、残りの 6,461,405 語 (5.2%) が未知語に由来する誤りだと考える。また、未知語全体のうち、語義タグ 1.1960 が付与されていたものは 27,177,289 語中の 7,238,358 語 (26.6%) に相当した。未知語かつ語義タグ 1.1960 が付与された単語のうち、頻出の単語の多くは全角の算用数字やダッシュ (—) のような記号である傾向が見られた。しかし、一部のレジスタでは他の単語が頻出であった。たとえば OL(法律) では「当該」、「前項」といった単語が、OM(議事録) では「ただいま」といった単語が頻出であった。

語義情報の悉皆付与に際して、学習データにおける未知語の WSD が今後の課題である。この解決のために、たとえば単語の品詞情報を明示的に利用した WSD システムにすることで、明らかなタグ付与の誤りを減らせると考えられる。

4　語義タグつき語彙表の構築

今回悉皆付与した語義タグつき BCCWJ に基づき、語義タグつき語彙表を構築した。前節に示した 1.1960 の語義タグについては、品詞が「名詞-数詞」ではない単語について解析困難語として x.xxxx の語義タグに修正を行った。

メタデータに基づいて 以下の 3 種類の語義タグつき語彙表を構築した。

- BCCWJ 全体の語彙表
- BCCWJ レジスタ毎の語彙表
- BCCWJ PN (新聞) メタデータつきの語彙表

表 3　BCCWJ 全体の語義タグの分布（粗頻度）

類　部門	異なり	（割合）	延べ	（割合）
1. 体	19,963	9.90%	34,497,391	27.79%
.1 関係	5,705	2.82%	16,523,255	13.31%
.2 主体	4,711	2.33%	5,982,156	4.82%
.3 活動	5,593	2.77%	8,237,250	6.63%
.4 生産物	2,045	1.01%	1,915,489	1.54%
.5 自然	1,909	0.94%	1,839,241	1.48%
2. 用	3,301	1.63%	13,790,441	11.11%
.1 関係	1,567	0.77%	6,330,852	5.10%
.3 活動	1,556	0.77%	7,305,325	5.88%
.5 自然	178	0.08%	154,264	0.12%
3. 相	2,736	1.35%	6,478,222	5.22%
.1 関係	1,676	0.83%	5,531,053	4.45%
.3 活動	802	0.39%	813,753	0.65%
.5 自然	258	0.12%	133,416	0.10%
4. 他	252	0.12%	1,255,310	1.01%
割当なし	175,521	87.05%	68,079,600	54.85%
（内 x.xxxx）	172,715	85.66%	6,461,405	5.20%
合計	201,626	100.00%	124,100,964	100.00%

表 3 に BCCWJ 全体の語義タグの分布を示す。類が体・相の語については、部門「.1 関係」が最も多く占めていた。類が用の語については、部門「.3 活動」の割合が多かった。

表 4 に BCCWJ レジスタ毎の上位頻度の語義タグを調整頻度 (pmw: per million word) とともに示す。「1.1960 体-関係-量-数記号（一二三)」、「2.1200 用-関係-存在-存在」、「2.3430 用-活動-行動-行動・活動」が頻度上位を占める傾向にあるが、レジスタによってその程度が異なることがわかる。さらに OL (法律)

表 4 BCCWJ レジスタ毎の頻度上位語義タグ (pmw)

レジ	順位	分類番号	ラベル	pmw
PB	1	2.3430	用-活動-行為-行為・活動	26146
書籍	2	1.1960	体-関係-量-数記号（一二三）	25738
	3	2.1200	用-関係-存在-存在	19025
PM	1	1.1960	体-関係-量-数記号（一二三）	48259
雑誌	2	2.3430	用-活動-行為-行為・活動	20361
	3	1.1962	体-関係-量-助数接辞	15164
PN	1	1.1960	体-関係-量-数記号（一二三）	54330
新聞	2	1.1962	体-関係-量-助数接辞	24699
	3	2.3430	用-活動-行為-行為・活動	23257
LB	1	2.3430	用-活動-行為-行為・活動	22896
書籍	2	1.1960	体-関係-量-数記号（一二三）	20119
	3	2.1200	用-関係-存在-存在	19870
OB	1	2.3430	用-活動-行為-行為・活動	20852
書籍	2	2.1200	用-関係-存在-存在	20187
	3	1.1960	体-関係-量-数記号（一二三）	13119
OC	1	2.3430	用-活動-行為-行為・活動	23801
知恵袋	2	2.1200	用-関係-存在-存在	18250
	3	1.1960	体-関係-量-数記号（一二三）	16417
OL	1	1.1960	体-関係-量-数記号（一二三）	58399
法律	2	2.3430	用-活動-行為-行為・活動	34383
	3	3.1101	相-関係-類-等級・系列	34257
OM	1	2.1200	用-関係-存在-存在	29570
国会	2	2.3430	用-活動-行為-行為・活動	29187
	3	2.3100	用-活 動-言 語-言 語 活動	22790
OP	1	1.1960	体-関係-量-数記号（一二三）	120276
広報紙	2	1.1962	体-関係-量-助数接辞	41483
	3	2.3430	用-活動-行為-行為・活動	18228
OT	1	1.1960	体-関係-量-数記号（一二三）	40084
教科書	2	2.3430	用-活動-行為-行為・活動	25478
	3	2.1200	用-関係-存在-存在	14446
OV	1	2.1200	用-関係-存在-存在	14094
韻文	2	2.3430	用-活動-行為-行為・活動	13906
	3	1.2010	体-主体-人間-われ・なれ・かれ	12651
OW	1	1.1960	体-関係-量-数記号（一二三）	68773
白書	2	2.3430	用-活動-行為-行為・活動	28421
	3	1.1962	体-関係-量-助数接辞	25557
OY	1	1.1960	体-関係-量-数記号（一二三）	34002
ブログ	2	2.3430	用-活動-行為-行為・活動	17627
	3	1.1962	体-関係-量-助数接辞	14348

に「3.1101 相-関係-類-等級・系列」が多く、OV (韻文) に「1.2010 体-主体-人間-われ・なれ・かれ」が多いなど、特徴的な傾向が確認された。

BCCWJ の PN (新聞) サンプルについては、新聞記事の境界情報とともに記事単位に分類情報が付与されている [10]。新聞記事の大分類情報ごとに語義タグの類と部門の粗頻度を係数したものを付録の表 5 に示す。この分割表に基づき、カイ二乗分析を行い、得られた標準化残差を付録の表 6 に示す。

新聞記事の語彙分布の差異は名詞（「1. 体」）に顕著にみられる。「1.1 体-関係」は経済・スポーツに多く、総類・文化に少ない。「1.2 体-主体」は政治・国際に多く、科学・社会に少ない。「1.3 体-活動」は政治・経済・国際に多く、科学・社会・スポーツに少ない。「1.4 体-生物」は経済・社会に多く、政治・スポーツに少ない。「1.5 体-自然」は科学・社会に多く、政治・スポーツに少ない。この体の部門の分布は、必ずしも用の部門の分布の傾向と相関しているわけではなく、例えば科学は「1.1 体-関係」が少ないが、「2.1 用-関係」が多い傾向がみられた。

5 おわりに

本研究では、BCCWJ を語義で検索できるようにすることを目標として、all-words WSD 技術を用いて『分類語彙表』の語義情報を付与した。言語研究において、BCCWJ のような大規模な言語データを語義に基づいて分析することで、言語の用法、意味を計量的に分析できるようになる。

本研究の有効性を検証するために語義情報つきの語彙表を構成し、レジスタごと・新聞記事分類ごとの分布の傾向を確認した。メタデータと結合することにより、語義の出現傾向の偏りを確認できた。また、学習データにおける未知語に対する語義情報付与が課題であることがわかった。

今後、検索系「中納言」への BCCWJ 語義検索機能の実装を進めるとともに、他のコーパスに対する語義情報付与を進める。

謝辞

本研究は JSPS 科研費 JP22K12145, 18K00634 及び国立国語研究所共同研究プロジェクト「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] 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 Resources and Evaluation**, Vol. 48, pp. 345–371, 2014.

[2] 国立国語研究所（編）. 分類語彙表増補改訂版. 大日本図書, 2004.

[3] Sachi Kato, Masayuki Asahara, and Makoto Yamazaki. Annotation of 'word list by semantic principles' labels for balanced corpus of contemporary written japanese. In **Proceedings of the 32nd Pacific Asia Conference on Language, Information, and Computation (PACLIC 32)**, 2018.

[4] 加藤祥, 浅原正幸, 山崎誠. 分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ. 日本語の研究, Vol. 15, No. 2, pp. 134–141, 2019.

[5] 近藤明日子, 田中牧郎. 「分類語彙表番号-UniDic 語彙素番号対応表」の構築. 国立国語研究所論集, No. 18, pp. 77–91, 2020.

[6] Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. All-words word sense disambiguation using concept embeddings. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[8] Soma Asada, Kanako Komiya, and Masayuki Asahara. All-words word sense disambiguation for historical japanese. In **The 37th Pacific Asia Conference on Language, Information and Computation (PACLIC 37)**, 12 2023.

[9] 国立国語研究所.『日本語歴史コーパス』, 2023. (バージョン 2023.3) https://clrd.ninjal.ac.jp/chj/.

[10] 加藤祥, 浅原正幸.『現代日本語書き言葉均衡コーパス』新聞サブコーパスの記事情報. 日本語の研究, Vol. 19, No. 2, pp. 206–214, 2023.

A 付録

表 5 BCCWJ PN (新聞) メタデータ付きの語彙表の語義タグの分布 （粗頻度）

	総類	政治	経済	労働	文化	科学	社会	事件	スポーツ	国際	見出し	総計
1.1 体-関係	22,811	36,445	35,747	5,054	29,940	14,650	49,080	21,757	40,610	17,449	632	274,175
1.2 体-主体	13,525	20,758	13,578	2,672	17,261	6,037	20,587	12,423	17,133	11,344	330	135,648
1.3 体-活動	13,886	25,205	19,485	3,055	17,752	6,863	22,107	11,956	16,539	12,611	389	149,848
1.4 体-生産物	2,327	1,785	3,629	400	2,912	1,453	5,020	2,179	1,986	1,114	52	22,857
1.5 体-自然	2,505	1,356	2,018	321	2,562	3,278	4,925	1,799	1,795	1,228	54	21,841
2.1 用-関係	6,586	8,313	7,097	1,384	8,004	4,237	10,764	5,193	8,067	4,459	102	64,206
2.3 用-活動	8,904	10,805	8,054	1,645	10,230	4,579	12,602	6,349	9,355	5,950	135	78,608
2.5 用-自然	284	133	110	37	278	149	332	103	195	62	4	1,687
3.1 相-関係	5,052	6,838	5,580	1,076	6,506	3,017	8,675	3,381	6,849	3,511	97	50,582
3.3 相-活動	1,040	994	710	152	1,220	451	1,527	510	1,149	509	15	8,277
3.5 相-自然	115	66	87	28	190	82	266	51	123	41	1	1,050
4. 他	800	665	521	132	739	306	1,142	358	524	359	16	5,562
タグなし	88,207	99,450	81,713	15,707	104,465	45,156	135,640	60,509	105,662	53,805	1,562	791,876
総計	166,042	212,813	178,329	31,663	202,059	90,258	272,667	126,568	209,987	112,442	3,389	1,606,217

表 6 BCCWJ PN (新聞) メタデータ付き語彙表の語義タグの分布に基づく標準化残差

	総類	政治	経済	労働	文化	科学	社会	事件	スポーツ	国際	見出し
1.1 体-関係	-38.10	0.73	35.43	-5.29	-28.78	-6.89	14.17	1.19	29.65	-14.34	2.45
1.2 体-主体	-4.64	23.31	-13.39	-0.04	1.68	-19.54	-18.44	18.26	-5.06	20.55	2.71
1.3 体-活動	-14.30	42.82	24.60	1.97	-8.99	-18.35	-24.07	1.49	-24.56	22.55	4.31
1.4 体-生産物	-0.78	-24.43	23.14	-2.42	0.74	4.88	20.23	9.34	-19.81	-12.69	0.55
1.5 体-自然	5.53	-30.90	-8.82	-5.37	-3.81	60.67	22.09	1.97	-21.43	-8.04	1.18
2.1 用-関係	-0.68	-2.30	-0.40	3.43	-0.89	11.00	-1.45	2.00	-3.91	-0.56	-2.94
2.3 用-活動	9.35	4.21	-7.84	2.51	3.76	2.57	-7.23	2.10	-10.00	6.41	-2.46
2.5 用-自然	8.77	-6.50	-5.99	0.66	4.83	5.73	2.96	-2.71	-1.85	-5.36	0.23
3.1 相-関係	-2.63	1.82	-0.52	2.56	1.95	3.43	1.06	-10.14	3.17	-0.53	-0.96
3.3 相-活動	6.67	-3.34	-7.33	-0.88	5.94	-0.68	3.58	-5.82	2.19	-3.04	-0.59
3.5 相-自然	0.65	-6.66	-2.91	1.62	5.39	3.08	7.22	-3.64	-1.31	-3.93	-0.82
4. 他	9.93	-2.85	-4.13	2.16	1.59	-0.38	7.08	-4.00	-8.09	-1.60	1.25
タグなし	32.90	-25.46	-31.17	1.10	23.07	4.51	5.10	-11.07	10.00	-10.08	-3.74