『現代日本語書き言葉均衡コーパス』長単位への分類語彙表番号悉皆付与

浅田宗磨　浅原正幸　柏野和佳子
国立国語研究所
{asada,masayu-a,waka}@ninjal.ac.jp

概要

本研究ではコーパスの検索性の向上のために、『現代日本語書き言葉均衡コーパス』（BCCWJ）に対して長単位への『分類語彙表』に基づく語義情報を悉皆付与した。BCCWJ-WLSP-LUW を学習データとした all-words WSD モデルを構築し、自動付与に用いた。学習に用いないテストデータによって性能を評価し、さらに付与結果に基づき語義分布をレジスタ別に分析した。

1　はじめに

本研究は、『現代日本語書き言葉均衡コーパス』（BCCWJ）[1] に対して長単位への『分類語彙表増補改訂版』[2] に基づく分類語彙表番号を悉皆付与し、語義情報とすることでコーパスの検索性を向上させることを目的とする。付与作業について、本研究ではエンコーダ型の言語モデルを用い、系列ラベリング手法に基づく all-words WSD 技術によって自動付与を行った。本稿では、利用したデータおよび解析手法を解説するとともに、BCCWJ の各レジスタに対し語義の分布の差異を検証する。また、BCCWJ に対する短単位における語義分布と長単位における語義分布の差異についても検証する。解析モデルは学習データを訓練データ・検証データ・テストデータに分割し、いくつかの言語モデルを試したうちテストデータで最良の結果を出したモデルとした。語義の分布について、分類語彙表の中項目を基にして調査した。

2　作業の内容

2.1　利用したデータ

『分類語彙表増補改訂版』[2] とは、国立国語研究所が作成した約 10 万語のエントリからなる語義情報付きの辞書である。語義情報は 5 桁からなる分類番号で表現される。本研究ではこれを語義タグとして扱う。ある単語に付与される分類番号は表 1 のように桁によって表示する意味が異なる。整数部分は類と呼び、主に品詞情報を表す。小数部分は意味分類であり、桁数が小さくなるにつれ、より詳細な分類（部門・中項目・分類項目の順）を表す。

表 1　例「会話」分類番号:1.3131

類	部門	中項目	分類項目
1	.3	.31	.3131
体	精神・行為	言語	話・談話

『現代日本語書き言葉均衡コーパス』（BCCWJ）[1] は 1 億語規模の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ベストセラー（OB）・Yahoo!知恵袋（OC）・法律（OL）・国会会議録（OM）・広報紙（OP）・教科書（OT）・韻文（OV）・白書（OW）・Yahoo!ブログ（OY）からなる 13 のレジスタで構成されている。全てのサンプルについて長短 2 つの言語単位を用いて形態素解析されている。短単位は現代語において意味を持つ最小単位で基本語彙の選定や用例収集を目的として規定される。長単位は文節を基にした単位で、語の使用実態の解明を目的として規定される。

BCCWJ-WLSP-LUW[3] は、PB・PM・PN の長単位に対し『分類語彙表増補改訂版』の分類語彙表番号が付与されたデータである。本研究では BCCWJ-WLSP-LUW を学習データとして用いた。

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。BCCWJ の短単位を対象にした all-words WSD の先行研究に浅田ら [4] の研究がある。この研究ではエンコーダ型言語モデルである BERT[5] を用

いて実験を行い、人手による評価によって BERT が all-words WSD に有効であることを示した。

本研究では先行研究 [4] に倣い、システムへの入力はタイトル情報を付与した文単位で行った。一文はコーパス中で句点（「。」）または文境界で区切られる箇所とした。入力文は言語モデルが用いるサブワード単位に分割（トークン化）したうえで、各トークンを語彙 ID（トークン ID）列に変換し、エンコーダに入力する。このとき、BCCWJ では一つのエントリとして扱われている語が、言語モデルのトークン化によって複数のトークンに分割される場合がある。たとえば「家庭訪問」は「家庭／訪問」のように 2 トークンに分割される。この場合、分割されたトークンはそのまま言語モデルに入力し、得られた各トークンの分散表現の平均を、元の語（長単位）の表現として用いる。

語義情報付与に用いるモデルの選定に際し、複数の言語モデルを WSD タスクに向けて fine-tuning し、最も性能の良かったモデルで付与を行うこととした。BCCWJ-WLSP-LUW を学習コーパスとして利用した。学習コーパスのうちから句点または文境界で区切られるトークン列を入力とした。学習コーパスを分割し、全体の 8/10 を訓練データに、1/10 を検証データに、1/10 をテストデータに充てた。学習コーパスを構築するにあたって、文の出現順序を無作為にシャッフルし、各分割データ中でレジスタに偏りがないようにした。学習は、入力データ列と正解語義列を用いて、入力データ列からモデルの出力する各トークンについての予測語義と正解語義を比較し一致するかどうかで進めた。学習に際して、あらかじめ BCCWJ-WLSP-LUW に登場する語について、登場する語義を収集してモデルが予測する時の候補として用いた。

性能の評価はテストデータで登場する多義語を対象語として正解率を算出し指標として用いることで行った。多義語は BCCWJ-WLSP-LUW に出現する語のうちで、複数種類の分類語彙表番号が付与されている語である。正解率は、式（1）のように算出した。

$$\text{正解率} = \frac{\text{対象語の正解数}}{\text{対象語の出現数}} \tag{1}$$

語義情報付与には、テストデータで最も性能の良かった、BERT を改善したモデルである ModernBERT[6] を基に作成された日本語モデルの Ruri[7] を用いた。また、Ruri を用いるにあたり、各文の先頭に「クエリ：」を付与して入力した。入力データ列中のタイトル情報及び「クエリ：」については「語義情報なし」のラベルを付与し、単義語として扱った。

3 評価

Ruri の最良モデルのテストデータでの正解率を表 2 に示す。全単語については、句点などの記号も含めるが、評価対象となる対象語中にタイトル情報および「クエリ：」のトークンは含まれない。

表 2 テストデータでの評価結果

対象語	正解	誤り	正解率 [%]
多義語	1,636	493	76.84
全単語	31,844	746	97.71

対象語を多義語とした時のモデルの正解率は 76.84%と、先行研究 [4] による短単位のものと比較すると 10 ポイント近く下がっていることが分かる。また、全単語の誤りのうち 70%程度を多義語が占めている。多義語での精度向上が今後の課題となる。

テストデータ中の正解例として、「関西国際空港（分類番号：1.2640「体-主体-社会-事務所・市場・駅など」）」を示す。この語は短単位では「関西（分類番号：1.2590「体-主体-公私-固有地名」）」、「国際（分類番号：1.3500「体-精神・行為-交わり-交わり」）」、「空港（分類番号：1.2640）」で構成される。このように、構成している短単位が複数の異なる分類番号を持っている場合でも正しく分類番号を付与できたことが分かる。また、別の例として「Q 3 具体的には、どういうときに不満を感じますか。」という文中の「具体的（分類番号：3.3070「相-精神・行為-心-意味・問題・趣旨など」）」という語を示す。この語は短単位では「具体（分類番号：1.1200「体-関係-存在-存在」）」と「的（分類番号：3.1130「相-関係-類-異同・類似」）」で構成される。長単位において、構成している短単位の分類番号と異なる語でも正しく分類番号を付与できたことが分かる。一方で、誤り例として、「複雑な感情を隠しつつ察してほしいと感傷的にならないほうがいい。」という文中の「感傷的（分類番号：3.3420「相-精神・行為-行為-人柄」）」を示す。この語は短単位では「感傷（分類番号：1.3014「体-精神・行為-心-苦悩・悲哀」）」と「的（分類番号：3.1130）」で構成される。モデルはこの語について「感傷的（分類番号：3.3300「相-精神・

行為-生活-文化・歴史・風俗」)」と誤って分類番号を付与していた。

長単位の語義情報を付与するに際して、より精度を向上させるために、短単位の語義情報を明示的に利用する方法が考えられる。また、そのうちでも意味分類を表す小数部分についてのみ利用することで、複合したときに品詞が変わっても対応して学習できると考えられる。

4 語彙の分布

4.1 各レジスタの比較

WSD システムによって長単位への分類語彙表番号を付与したことで、BCCWJ の各レジスタについて分布の比較が行えるようになった。表 3 に BCCWJ レジスタごとの上位頻度の中項目までの分類番号を調整頻度（pmw：per million word）とともに示す。本稿の以降の「頻度」はこの調整頻度を表す。

各レジスタの頻出中項目を見ると、1.19「体-関係-量」、2.12「用-関係-存在」、2.15「用-関係-作用」、2.30「用-精神・行為-心」が多くのレジスタで上位となっていることが分かる。OL については、他のどのレジスタでも上位とならなかった 4.11「他-接続」、1.10「体-関係-事柄」、1.30「体-精神・行為-心」が上位となっている。OP については、1.19 が群を抜いて多く登場し、それ以降の順位のものも 1.16「体-関係-時間」、1.24「体-主体-成員」と他では頻出上位でないものがあがっている。

図 1 に各レジスタの中項目の頻度を用いた t-SNE（perplexity=3, seed=42）による 2 次元プロットを示す。この図から、各レジスタの分布について大きく分けて 3 つのグループが存在することが分かる。左下には OB、OC、OY が集まってプロットされている。中央には LB、OM、OV、PB、PM が集まってプロットされている。右上には OL、OP、OT、OW、PN が集まってプロットされている。これらのグループ分けは、各レジスタでの頻出の中項目の順位が同じだから同グループに所属するというものではないことが見てとれる。

4.2 短単位との比較

図 2 に長単位と短単位の中項目の分布を示す。グラフの線が切れている箇所は該当する中項目が出現しなかったことを表す。長単位は短単位と比較して語義情報なしの語が多く、全体的に頻度が小さめと

表 3 各レジスタでの頻出中項目

レジスタ	順位	番号	ラベル	pmw
LB（書籍）	1	1.19	体-関係-量	6,239
	2	2.12	用-関係-存在	6,122
	3	2.15	用-関係-作用	5,545
OB（ベストセラー）	1	1.19	体-関係-量	5,388
	2	2.12	用-関係-存在	5,163
	3	2.15	用-関係-作用	4,771
OC（知恵袋）	1	1.19	体-関係-量	6,092
	2	2.30	用-精神・行為-心	5,028
	3	2.12	用-関係-存在	4,446
OL（法律）	1	4.11	他-接続	11,068
	2	1.10	体-関係-事柄	8,490
	3	1.30	体-精神・行為-心	7,195
OM（国会）	1	2.12	用-関係-存在	6,973
	2	1.19	体-関係-量	5,319
	3	2.30	用-精神・行為-心	4,903
OP（広報紙）	1	1.19	体-関係-量	45,033
	2	1.16	体-関係-時間	3,981
	3	1.24	体-主体-成員	3,703
OT（教科書）	1	1.19	体-関係-量	10,613
	2	2.15	用-関係-作用	6,061
	3	2.30	用-精神・行為-心	5,455
OV（韻文）	1	2.12	用-関係-存在	6,723
	2	2.15	用-関係-作用	5,563
	3	1.19	体-関係-量	4,978
OW（白書）	1	1.19	体-関係-量	12,563
	2	2.15	用-関係-作用	8,413
	3	2.30	用-精神・行為-心	4,610
OY（ブログ）	1	1.19	体-関係-量	8,397
	2	2.15	用-関係-作用	3,864
	3	2.30	用-精神・行為-心	3,851
PB（書籍）	1	1.19	体-関係-量	6,512
	2	2.12	用-関係-存在	5,862
	3	2.15	用-関係-作用	5,840
PM（雑誌）	1	1.19	体-関係-量	9,397
	2	2.15	用-関係-作用	4,907
	3	2.30	用-精神・行為-心	4,344
PN（新聞）	1	1.19	体-関係-量	16,362
	2	2.15	用-関係-作用	5,332
	3	2.30	用-精神・行為-心	4,706

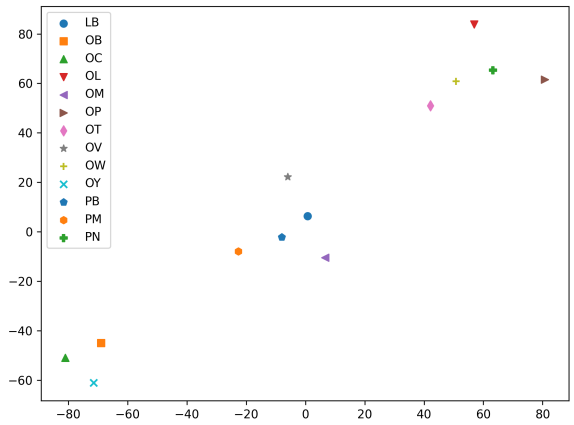

図 1　各レジスタでの中項目の分布の t-SNE

なっている。そのため体の類では多くの項目で 10 倍程度の頻度の開きがあるが、その分布傾向は似ている。どちらも 1.19「体-関係-量」が最頻出である。

用の類では多くの項目で頻度が数倍の開きにとどまり、2.12「用-関係-存在」、2.15「用-関係-作用」、2.30「用-精神・行為-心」がどちらも頻出である。2.34「用-精神・行為-行為」については長単位と短単位で出現頻度に 28 倍以上の開きがある。これは、2.3430「用-精神・行為-行為-行為・活動」に「する」という語が存在しているからだと考えられる。この語はたとえば「勉強する」のように名詞と接続して複合動詞を作るが、短単位では「勉強／する」のように 2 つの語として、長単位では 1 つの語として扱われる。全体的に見ると、長単位は用の類が短単位と比較して多く出現していることも同様の理由からであると考えられる。2.17「用-関係-空間」については長単位での出現頻度が短単位を上回っている。

相の類については、長単位では 3.12「相-関係-存在」、3.30「相-精神・行為-心」が、短単位では 3.19「相-関係-量」、3.10「相-関係-真偽」が頻出である。3.17「相-関係-空間」については長単位での出現頻度が短単位を上回っている。

5　おわりに

本研究では、コーパスの検索性の向上のために、BCCWJ の長単位への『分類語彙表増補改訂版』に基づく語義情報の付与を all-words WSD 手法によって行った。長単位の語義情報の分析が可能になると、言語的特徴の解明の一助となる。

本研究で付与した語義情報の分布をレジスタ毎または長短単位で比較した。出現頻度の高い語義は多くのレジスタで共通であった。短単位と長単位では各類における分布傾向の差異を見ることができた。

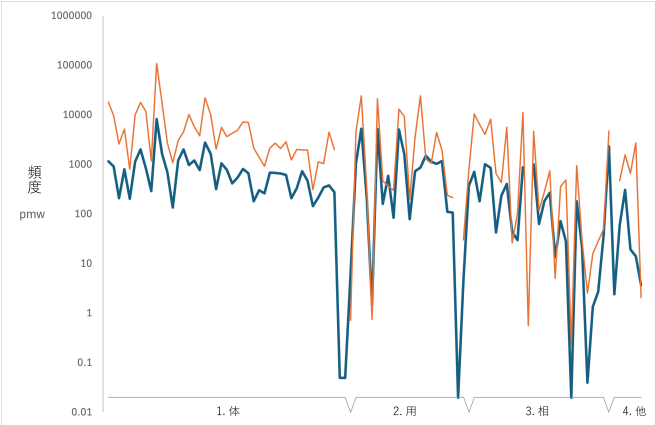

図 2　中項目の分布（細橙：短単位、太青：長単位）

今後、本研究で付与した語義情報の有用性を人手による検証によって示したい。

謝辞

本研究は、国立国語研究所共同研究プロジェクト「多様な語彙資源を統合した研究活用基盤の共創」「実証的な理論・対照言語学の推進」によるものです。

参考文献

[1] 山崎誠（編）.『書き言葉コーパス—設計と構築—』. 講座日本語コーパス 2. 朝倉書店., 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] 加藤祥, 浅原正幸. Bccwj-wlsp-luw:『現代日本語書き言葉均衡コーパス』に対する長単位語義情報アノテーション. 言語処理学会, 2025.

[4] 浅田宗磨, 古宮嘉那子, 浅原正幸.『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. 言語処理学会, 2024.

[5] 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, 2019.

[6] Bonaventure Chidube Molokwu, Audrey Rah, and Reginald Chukwuka Molokwu. Fine-grained sentiment mining, at document level on big data, using a state-of-the-art representation-based transformer: Modernbert. 2025.

[7] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings, 2024.