BERT を使用した日本語の単語の通時的な意味変化の分析

小林千真 [1] 相田太一 [2] 小町守 [2]
法政大学 [1] 東京都立大学 [2]
kazuma.kobayashi.3t@stu.hosei.ac.jp {aida-taichi@ed.,komachi@}tmu.ac.jp

1 はじめに

時間の経過で単語の使われ方や意味は変化することがある。例えば、「こだわる」という単語は最近では広告でもポジティブな意味で使われているが、元々の意味は必要以上に気にするというネガティブな意味であった[1)]。このような現象は日本語に固有のものではなく、その他の言語でも確認されている [1]。単語の意味変化を定量的に分析する技術は言語学や辞書学における単語の意味変化の分析の助けとなることが期待される。

単語の意味変化を捉える研究として、単語分散表現を用いる手法が提案されている [2, 3, 4]。しかし、これらはどれも各期間で 1 単語につき 1 つのベクトルを対応させ、それらを比較することで単語の意味変化を検出している。また、時武ら [5] は単語の意味の広がりを捉えるために、ガウス埋め込みを用いて通時的な単語の意味変化を分析する手法を提案した。しかし、これらの手法は意味の検出はできるが、単語が持つ複数の意味がどのように変化しているのかを捉えることは難しい。これは 1 単語タイプに 1 つの表現を与えるためである。

これに対して、BERT [6] に代表される事前学習済み言語モデルを使えば各単語トークンが出現する文脈に応じた単語ベクトルを獲得することができるため、単語がどの意味で使用されているのかを単語トークンレベルで捉えることができる。近年の研究では、Hu ら [7] の BERT から得た単語ベクトルを辞書に載っている語義に対応づけをする研究、Giulianelli ら [8] の BERT から得た単語ベクトルの集合にクラスタリングを行う研究などが行われている。しかし、どちらの研究も英語でのみ行われている。

日本語における単語の意味の通時的変化を捉える研究としては相田らの研究 [9] がある。しかし、彼らの研究でも先程述べたように 1 単語タイプにつき 1 つのベクトルを対応させているため、1 単語に複数の語義を含む場合、それらを別々に扱うことができない。

本研究では、日本語を対象として Hu らの手法と Giulianelli らの手法をそれぞれ適用し、その結果を比較・分析した。本研究での貢献を以下に示す。

- 日本語において、BERT を使用して単語の意味の通時的変化の分析をした。
- Hu らの辞書を使用する手法と Giulianelli らのクラスタリングを使用する手法を比較した。

2 BERT を用いた通時的な語義分類

本研究では、日本語に対して Hu らの手法と Giulianelli らの手法を用いて実験し、比較した。

通時的なコーパスを一文ずつ BERT に入力し、対象となる単語の単語ベクトルを獲得する。以下では、獲得した単語ベクトルの集合を S とする。

2.1 辞書を用いた語義分類

まず、Hu らの辞書を用いた手法を説明する。n 個の語義を持つ単語に対して、辞書に載っている語義に対応する例文を BERT に入力し、語義ベクトル $v_0, v_1, ...v_n$ を獲得する。ただし、1 つの語義に対して複数の例文が載っている場合は、獲得した単語ベクトルの平均を語義ベクトルとする。v_x は各語義ベクトルであり、S の各要素は v_x とのコサイン類似度が最も高い語義に属するものとする。

2.2 クラスタリングを用いた語義分類

次に、Giulianelli らのクラスタリングを用いた手法を説明する。S に対してクラスタ数を 2 から 10 まで k-means 法を使ってクラスタリングを行い、silhouette score [10] の平均が最大のクラスタ数におけるクラスタリングの結果を採用することで、語義数と単語ベクトルが属するクラスタを決定する。ただし、k-means 法における距離はユークリッド距離

1） https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/kokugo_yoronchosa/pdf/92701201_04.pdf

を使用する。

3 実験設定

3.1 データセット

幅広い年代にわたる日本語コーパスとして、国立国語研究所の『日本語歴史コーパス』[2] の一部として公開されている近代雑誌コーパス に、「昭和・平成書き言葉コーパス 」[3] として構築中の雑誌（『中央公論』『文藝春秋』）データを追加したものを用いた。コーパス全体は 1874 年から 1997 年までに刊行された雑誌から成るが、これを 1898 年から 1997 年の 100 年間を 25 年区切りで 4 分割したものを使用した。

3.2 対象とする単語

一般的に意味の変化が起きていると知られている日本語の単語として、「こだわり」、「敷居[4]」、「適当[5]」、「全然[6]」などがある。本研究ではこれらの単語を対象とした。ただし、動詞の「こだわる」の活用形としての「こだわり」は対象とせず、名詞の「こだわり」のみを対象とした。また、前処理として対象の単語における表層形の正規化を行った。

3.3 辞書

goo 辞書[7] の国語辞典（デジタル大辞泉。2020 年 12 月時点。）と日本国語大辞典（第二版）を使用した。基本的に、3.2 節で示した単語の goo 辞書に載っている意味と例文（全て）を使用したが、「敷居」は例文が記載されていなかったため、日本国語大辞典を使用した。ただし、日本国語大辞典の例文は、もっとも新しい年代の例文を 1 つ使用した。

3.4 実験方法

1898 年から 1997 年のコーパスの全文を BERT に入力し、対象の単語の単語ベクトルを獲得する。ただし、本研究では、事前学習済みモデルとして、huggingface[8] で公開されている東北大の日本語版 BERT[9] を使用した。このようにして得た単語ベクトルの集合に対し、各対象単語で辞書とクラスタリングの手法をそれぞれ適用し、単語ベクトルの集合をグルーピングした。期間ごとにその結果を語義の使用比率を示した積み上げ棒グラフで表した。ここで、辞書の手法の凡例には辞書の語義をそのまま示し、クラスタリングの手法の凡例には各クラスタ内で主に確認された用例を示した。また、単語ベクトルとそれに対応する実際の用例を比較することで分析をした。

4 結果と考察

図 1 に辞書の手法とクラスタリングの手法それぞれの実験で得られた各単語の通時的な語義の使用比率の変化を示す。また、図 2 に「全然」における各手法の単語ベクトルの分類の結果と期間ごとの単語ベクトルの分布を主成分分析で可視化した散布図を示す。

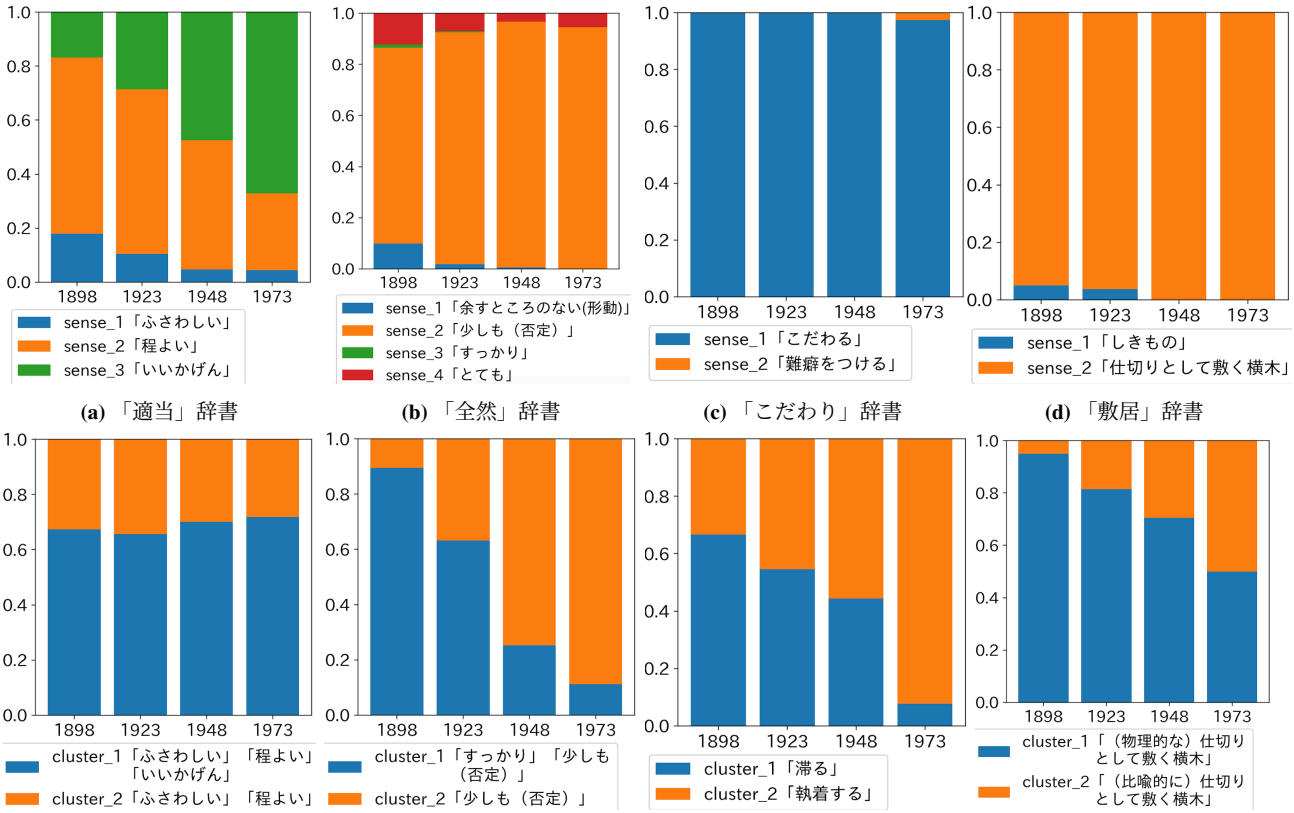
4.1 事例分析

「適当」 図 1a の『辞書』では sense_3 の「いいかげん」の割合が徐々に大きくなっており、3.2 節で想定していた結果を得た。図 1e の『クラスタリング』の手法では、通時的な変化は捉えられなかった。

実際に単語ベクトルに対応する文を確認すると、どちらの手法においても、「ふさわしい」の用例が全てのクラスタに出現していた。具体的には、『辞書』において、sense_1 では「ふさわしい」の用例がほとんどであり、sense_2 には「ふさわしい」と「程よい」の用例が多く属していた。sense_3 には「ふさわしい」と「いいかげん」の用例が多く見受けられた。『クラスタリング』において、cluster_1 には「ふさわしい」、「程よい」、「いいかげん」の 3 つの用例があり、cluster_2 では「ふさわしい」と「程よい」の用例があった。

「全然」 図 1b の『辞書』では、ほとんどの単語ベクトルが sense_2 の「少しも（否定）」に属してしまい、通時的な変化を捉えられなかった。図 1f の『クラスタリング』では「すっかり」と「少しも（否定）」の用例を含んでいる cluster_1 の割合が徐々に小さくなっている。「すっかり」の用例は cluster_1 でのみ出現しており、3.2 節で想定した結果が得ら

2） https://pj.ninjal.ac.jp/corpus_center/chj/
3） https://pj.ninjal.ac.jp/corpus_center/cmj/woman-mag/
4） 「敷居」という単語には「敷居が高い」という熟語があり、「不義理をして行きにくい」という意味から「気軽にはいけない」という意味に変化している。
5） 「適当」は「ふさわしい」と「程よい」という意味に加え、近年では「いいかげん」という意味でも使われている。
6） 「全然」は明治時代から戦前までは「すっかり」という意味で使用されることが多かったが、戦後は「（否定の表現を伴って）少しも」という使い方が一般的になった。
7） https://dictionary.goo.ne.jp/
8） https://github.com/huggingface/transformers
9） https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/



(a) 「適当」辞書　**(b)** 「全然」辞書　**(c)** 「こだわり」辞書　**(d)** 「敷居」辞書

(e) 「適当」クラスタリング　**(f)** 「全然」クラスタリング　**(g)** 「こだわり」クラスタリング　**(h)** 「敷居」クラスタリング

図 1: 通時的な語義の使用比率を表した棒グラフ。

れた。

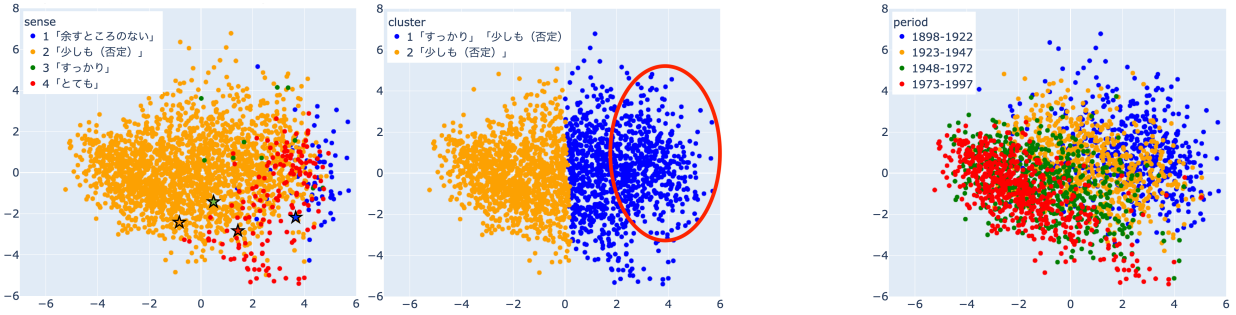
実際に単語ベクトルに対応する文を確認すると、『辞書』において sense_1 には形容動詞の「余すところのないさま」の意味ではなく、「すっかり」の用例が属していた。sense_2 には「少しも（否定）」と「すっかり」の用例が多く属していた。sense_3 は「すっかり」の用例が属していた。sense_4 には「とても」ではなく、「少しも（否定）」と「すっかり」の用例が見受けられた。コーパスには「すっかり」の用例は sense_3 に属する単語以外にも多く存在していたが、sense_3 にはわずかしか属さなかった。また、コーパスには「とても」と「余す所のない（形動）」の用例はほとんどなかった。『クラスタリング』において、cluster_1 には「すっかり」と「少しも（否定）」の 2 つの用例が多く見られ、cluster_2 では、ほとんどが「少しも（否定）」の意味の用例だった。

主成分分析で「全然」の単語ベクトルを可視化した図 2a,2b に注目すると、それぞれの手法の分類結果は様子が異なることが分かった。これは辞書を使った手法では例文から構築した語義ベクトル（図 2a の星）とのコサイン類似度に注目しているのに対して、クラスタリングでは k-means 法を使用し、セントロイドからの距離に注目していることによる違いである。ここで、図 2b の赤線で囲った部分に「すっかり」の用法がまとまっていた。また、図 2c に示した期間ごとの分布から年代が新しくなるにつれて徐々に、単語ベクトルの分布が右上から左下にシフトしていくことを確認した。

「こだわり」　図 1c の『辞書』ではほとんどが sense_1 に属しており、意味変化を捉えていなかった。図 1g の『クラスタリング』では、「滞る」[10] の用例が属してた cluster_1 の割合が徐々に減少しており、3.2 節で想定していた結果とは異なるが、意味の通時的変化を捉えた。

実際に単語ベクトルに対応する文を確認すると、『辞書』において、sense_1 は「執着する」と「滞る」の用例、sense_2 は「執着する」の用例が 1 件のみ属していた。コーパスには「難癖をつける」の用例は存在していなかったためこのような結果になったと考える。『クラスタリング』において、cluster_1 は「滞る」の用例、cluster_2 は「執着する」の用例をそれぞれ多く含んだ。

10)　コーパスの例文：「疾患のほかに避妊手術の後遺症で精神障害が起きたことを、二人はこだわりなく話した。」



(a)「全然」辞書　(b)「全然」クラスタリング　(c)「全然」期間ごとの分布

図 2: 主成分分析を使用し、「全然」の単語ベクトルを可視化した散布図。

「敷居」　図 1d の『辞書』では、ほとんどが sense_2 に属しており、意味変化を捉えていなかった。図 1h の『クラスタリング』では、比喩的な用法が属する cluster_2 の割合が徐々に大きくなり、3.2 節で想定していた意味変化とは異なるが、意味の通時的変化を捉えた。

実際に単語ベクトルに対応する文を確認すると、『辞書』において、sense_1 は「仕切りとして敷く横木」の用例が属していたが 2 件のみであり、sense_2 は「仕切りとして敷く横木」の用例が属していた。辞書に載っている「敷居」の意味は「しきもの」と「仕切りとして敷く横木」であるが、コーパスでは前者はほとんど出現しないため、sense_1 に分類されたものがほとんどなかったと考える。『クラスタリング』において、cluster_1 では物理的な「仕切りとして敷く横木」の用例、cluster_2 では「... 敷居が高い..」や「... 血の敷居は越え...」のように物理的な意味から拡張し抽象化された用例が多く出現していた。

4.2　考察

辞書を使用した手法では、辞書に載っている語義に対応した用例を捉えることが多かったが、「全然」では語義の違いを捉えられなかった。一方で、クラスタリングを使用した手法では、辞書には載っていない抽象的な意味を捉えることや辞書に載っているよりも詳細な違いを捉えることがあるとわかった。このため、クラスタリングを使用した手法は人間が意識していない意味の変化を捉えることを期待できるといえる。

「適当」における辞書の手法と「全然」におけるクラスタリングの手法ではそれぞれ、結果としては意味の通時的な変化を捉えたが、各クラスタが単一の用例で構成されていないため、適切な分類ができていない可能性があると考えた。

図 2c における年代による分布のシフトは「全然」の意味が徐々に変化する様子を捉えており、分類が難しかったと予想される。また、辞書の手法で、全てのクラスタに「すっかり」が含まれていたのは、語義ベクトルが古い時代の分布に近い位置にあったことや BERT が通時的なコーパスに適応できなかったことが原因と考える。

その他にも辞書の手法では、例文の数が十分でなかったことや辞書の語義が網羅されていない場合があることが通時的な意味変化を捉えられなかった原因と考える。クラスタリングの手法では、k-means 法がそれぞれのクラスタがだいたい同じ大きさであることを仮定するため、語義の頻度に大きな偏りがある場合は適切でないことがあると考える。

5　おわりに

本研究では BERT を使用して単語の通時的な意味変化を追う 2 つの手法を日本語に適応し、比較することでそれぞれの手法の特徴を確認した。辞書の手法では、辞書に載っている意味しか扱えないため、目的の意味を扱えないことや辞書の例文をそのまま使うのでは不十分である可能性があることを確認した。クラスタリングの手法では、辞書に載っていない変化を捉えることが期待できることや k-means 法が各クラスタが大体同じ大きさであることを仮定するので単語の意味を分類しきれない事例があることを確認した。

また、そもそも BERT から得た単語ベクトルが意味の違いを捉えきれない可能性があるため、今後の研究として、BERT から得られる単語ベクトルの特徴について調査を行いたい。

謝辞　本研究は JSPS 科研費 19H00531 の助成を受けたものである。

参考文献

[1] 岩本蘭, 湯川正裕. Auto-antonym の多カーネルオンライン意味変化分析. 言語処理学会第 25 回年次大会発表論文集, pp. 171–174, 2019.

[2] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 61–65, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

[3] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, p. 673–681, New York, NY, USA, 2018. Association for Computing Machinery.

[5] 時武孝介, 村脇有吾, 黒橋禎夫. ガウス埋め込みに基づく単語の意味の史的変化分析. 言語処理学会第 24 回年次大会 発表論文集, pp. 61–64, 2018.

[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] Renfen Hu, Shen Li, and Shichen Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3899–3908, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3960–3973, Online, July 2020. Association for Computational Linguistics.

[9] 相田太一, 小町守, 小木曽智信, 高村大也, 坂田綾香, 小山慎介, 持橋大地. 単語分散表現の結合学習による単語の意味の通時的変化の分析. 言語処理学会第 26 回年次大会 発表論文集, pp. 485–488, 2020.

[10] Peter J.Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. In *Journal of Computational and Applied Mathematics, Volume 20*, pp. 53–65, 1987.