日本語 BERT を用いた単語の用例の分野別分析ツールの開発

凌 志棟　相田 太一　金 輝燦　岡 照晃　小林 千真　小町 守
東京都立大学
{ling-zhidong,aida-taichi,kim-hwichan,kobayashi-kazuma1}@ed.tmu.ac.jp
{teruaki-oka,komachi}@tmu.ac.jp

概要

「保守（maintenance vs conservative）」のように同じ字面でも媒体によって意味が変わる単語がある。こうした単語の使用実態を調査する場合、用例を集め、分野別に分け、最後に分析を実施する。しかし大規模な分析調査となると、必要な時間や人件費が膨大になる。そのため検索システムや分析手法の開発など、計算機による研究支援の発展が期待されている。本研究では、日本語を対象とし、文脈を考慮した単語ベクトルで用例をクラスタリング、プロットする可視化ツールを作成した。単語が入力されると、現代日本語書き言葉均衡コーパスでの用例を集めてクラスタリングし、そのクラスタとコーパスに付与されている分野情報を合わせて 2 次元にプロットする。コーパス中の単語はすべて NWJC-BERT でベクトル化してあり、クラスタリングとプロットに使用される。また各プロット点からその用例を見ることも可能である。最後にこの可視化ツールが用例分析にどのように役立つのか、ケーススタディを紹介する。

1　はじめに

コーパス使った大規模な用例分析が言語学の研究で盛んになっている。久屋 [1] は外来語「ケース」を対象に、書き手の生年代、学歴や媒体が用法の違いに与える影響について統計分析を行なっている。その結果、前述の要因が外来語の生起に影響をもたらすことがわかった。また呉 [2] は「足を洗う」という表現の用例を分析した。その結果、分野特有の文脈語と共起することで意味の異なりが起きる傾向を明らかにした。久屋 [1] や呉 [2] のように分野間で用例を比較・分析することで、共時的分布を明らかにし、その単語の意味や用法に違いが生まれる要因を特定することができる。こうした研究は従来人手で行われてきたため、計算機を用いたコーパス分析への期待は大きい。

国立国語研究所が公開しているコーパス検索システム「中納言[1)]」を用いることで、現代語に限らず、古文、方言、学習者の書いた文の大規模な用例検索と収集は可能である。しかし中納言は検索機能しか提供していないため、用例に対する分析や統計的処理は研究者自身が人手で行う必要がある。このため実際に用例を統計的に分析し終えるまで対象単語が分析の対象として適切かどうかは判断できない。

そこで本研究では、調査対象となる単語（以下、**対象単語**）の分野別用例調査を効率化する可視化ツールを開発した。このツールを使うことで、対象単語の分野間で起こる用法の差異をより視覚的に捉えることができる。具体的には、幅広い分野をカバーした現代日本語書き言葉均衡コーパス（以下、BCCWJ）[3] の各文に対し、BERT [4] の単語ベクトルを用いた語義変化検出の手法 [5] を適用することで、単語の用例の違いの可視化を実現した。このツールによって、用例分析前に対象単語が研究対象としてそもそも適切かどうか判断が可能になる。本ツールは Google Colaboratory の形式で公開している[2)]。

2　関連研究

2.1　BERT を用いた語義変化検出

Giulianelli ら [5] は文脈を考慮した単語ベクトルを用い、単語の通時的な教師なし語義変化検出手法を提案した。年代の異なるコーパスから対象単語の用例を抽出し、BERT に 1 文ずつ入力し、対象単語のベクトルを獲得する。これらのベクトルに対しクラスタリングを行い、類似する用例ごとにまとめあげる。各文に付与された年代情報を用い、時間順に用

1）https://chunagon.ninjal.ac.jp/
2）https://colab.research.google.com/drive/1M1L531l2oxvX3pGkCzYEQmuBw_mXX9Ee

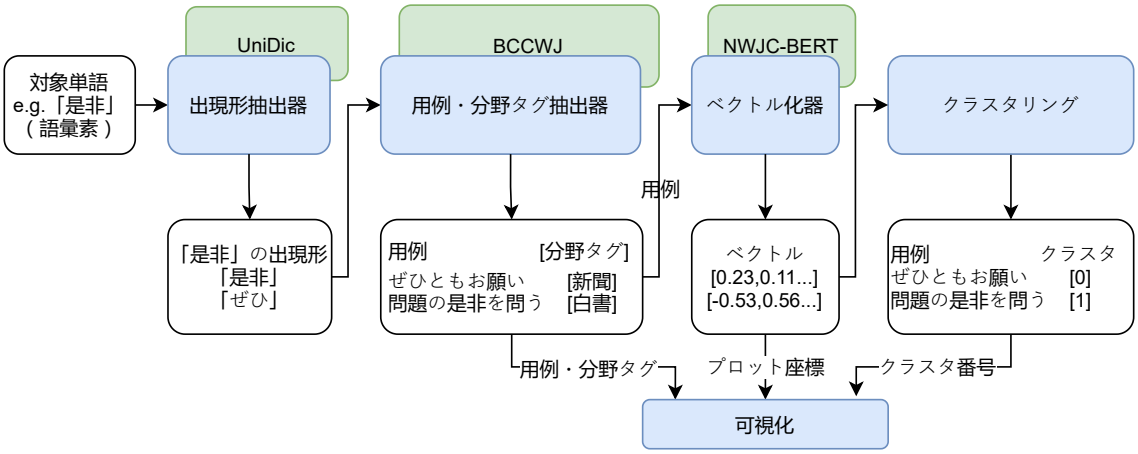


図 1: 開発したツールの処理フロー。「是非」を例にした場合の可視化の流れ。

例をプロットする。年代ごとのクラスタの割合の推移から、語義の通時変化を検出する。Giulianelli ら [5] の手法は通時的な変化の検出に向けたものであるが、本研究ではそれを異なる分野間での単語の用例分析（共時的な差異の分析）に適用した。また BCCWJ 中の文[3] には、その文が記述されていた文献のレジスター（ジャンル、言語使用域）が付与されている。本稿ではこれを**分野**情報と呼ぶ。クラスタリングの結果と、BCCWJ 分野情報を合わせて可視化することで、単語の用法（用例クラスタ）と分野との関係を発見することが可能となる。

2.2 BCCWJ の分野情報

BCCWJ は現代日本語の書き言葉の全体像を把握するために構築された日本語均衡コーパスで、全体で 1 億 430 万語規模あり、うち人手アノテーションされたコアデータは 100 万語の規模を持つ。実社会のさまざまなテキストを収集しており、各テキストには収集元の分野情報が付与されている。BCCWJ に収録されている分野と語数（短単位数）の一覧を参考情報内の表 1 に示す。本稿で紹介する可視化ツールでは、対象単語の用例を 2 次元上にプロットするだけでなく、その用例が出現した分野の情報を確認することができる。

3 対象単語の用例分析に向けた可視化ツールの構築

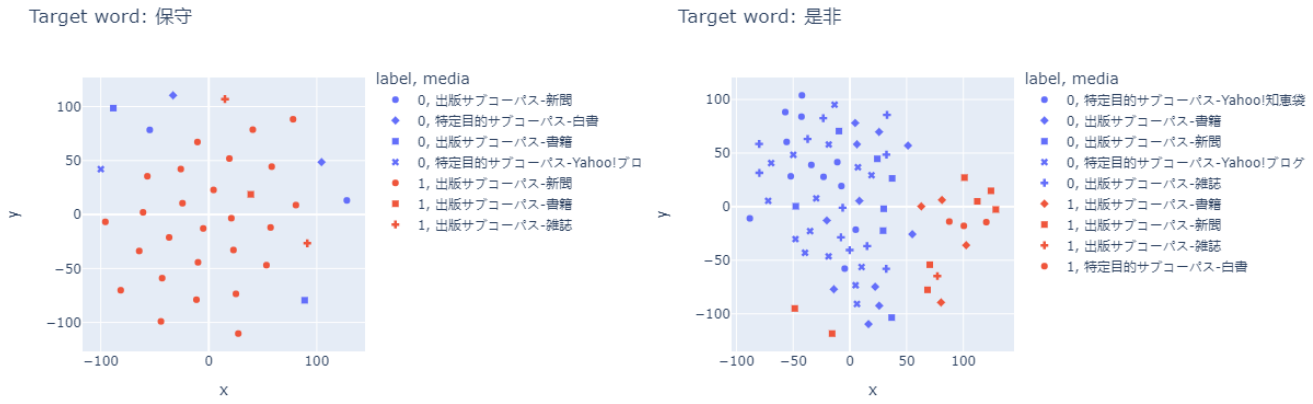
本研究では、BERT のベクトルを利用して、単語の用法の違いを分野別に可視化するツールを作成した。本ツールの処理の概要を図 1 に示す。対象単語を決め、語彙素（辞書の見出し語に相当するもの）[6][4] の形式でツールに入力する。出現形抽出器は単語の電子化辞書 UniDic を用い、同一語彙素を持つ出現形を網羅的に抽出する。UniDic は階層的見出し構造を採用しており、語彙素の下に語形、語形の下に出現形が保持されている。そのため例えば、動詞「食べる」という語彙素で検索する場合、出現形抽出器は同一語彙素「食べる」を持つ「食べ」、「食べよ」、「たべ」... を抽出する。抽出した出現形すべてで BCCWJ 内の文を検索し、それらを含む文を分野情報とともに抽出する（重複排除）。抽出された文を 1 文ずつ BERT に入力し、対象単語のベクトルを獲得する。ベクトルを Kmeans 法でクラスタリングし、対応する用例、クラスタ番号、BCCWJ の分野を付与した上で、2 次元に描画する。

BERT には NWJC-BERT[5] を使用する。NWJC-BERT は「国語研日本語ウェブコーパス」（NWJC）により事前学習した BERT モデルである。広く利用されている東北大 BERT[6] と異なり、NWJC-BERT は単語をサブワード化しない。対象単語がサブワード化されると、単語がより小さいトークンに分割され、単語ベクトルを獲得するためには、各トークンのベクトルを再度、ベクトル演算によって結合する必要がある。一方 NWJC-BERT では、語彙に UniDic の語彙素を収録しているため、対象単語が必要以上に分割される心配がない。したがって、NWJC-BERT を利用することでツールの処理が容易になる。

クラスタリング手法は Giulianelli らの研究と同じく Kmeans [7] を使用する。単語ベクトル集合に対し

3） 本稿では BCCWJ 中の文を「用例」として扱う。

4） BCCWJ, UniDic で lemma と呼ばれる要素。
5） https://www.gsk.or.jp/catalog/gsk2020-e/
6） https://github.com/cl-tohoku/bert-japanese
7） https://scikit-learn.org/stable/modules/generated/



(a) 色でクラスタを区別した「保守」のグラフ　　**(b)** 色でクラスタを区別した「是非」のグラフ

図 2: 作成したツールで各対象単語に対して用例ごとにクラスタリングを行った結果

プロット点	クラスタ番号(色)	分野タグ
●	0(青)	特定目的サブコーパス-Yahoo!知恵袋
◆	0(青)	出版サブコーパス-書籍
■	0(青)	出版サブコーパス-新聞
✖	0(青)	特定目的サブコーパス-Yahoo!ブログ
✚	0(青)	出版サブコーパス-雑誌
◆	1(赤)	出版サブコーパス-書籍
■	1(赤)	出版サブコーパス-新聞

図 3: 可視化ツールの凡例の見方

て、クラスタ数を 2 から 10 まで Kmeans を行い、シルエットスコアが最大となるクラスタ数での結果を採用する。類似度計算にはユークリッド距離を使用する。

4　可視化ツールを用いたケーススタディ

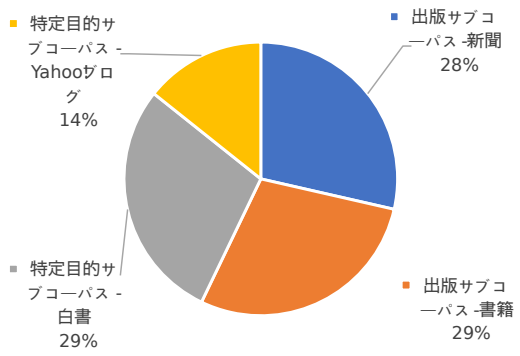
この可視化ツールを用いて単語「保守」と「是非」に対し分析を行なった。用例抽出の対象は BCCWJ コアデータを使用した。対象単語のベクトルを可視化する際に、scikit-learn の TSNE [8] を用いて 768 次元のベクトルを 2 次元に次元圧縮を行なった。図 2a、2b に単語「保守」と「是非」に対して本ツールを用いた可視化結果を示す。出力された図の凡例の見方は図 3 に示す。

ケーススタディ 1「保守」　図 2a に「保守」という単語の用例に対するクラスタリングの結果を示す。コトバンク辞書と見比べた時、Kmeans によるクラスタリングの結果はクラスタ 0（青）が「正常の状態を保つ」[7] という意味での用例、クラスタ 1（赤）が「旧習・伝統を守る」[7] という意味の用例であった。また各用例には BCCWJ 由来の分野情報が付与されている。クラスタ 1（赤）「旧習・伝統を守る」に含まれる用例の出典分野の割合を図 4a に示す。これを見ると「旧習・伝統を守る」という意味での「保守」は新聞分野での用例が全体のおよそ 90 %を占めることがわかる。これに対し図 4b にはクラスタ 0（赤）「正常の状態を保つ」に含まれる用例の出典割合を示した。この意味での使用は新聞やブログなどの複数の分野で出現した上、それぞれの分野でほぼ同じ割合を占めている。そのため「保守」という単語は「正常の状態を保つ」という意味で幅広く使われていることが分かった。プロットからは直接用例を確認できるため、用例文を詳しく見ていくと「保守派」「保守党」という政治関連での出現が多いことが分かった。

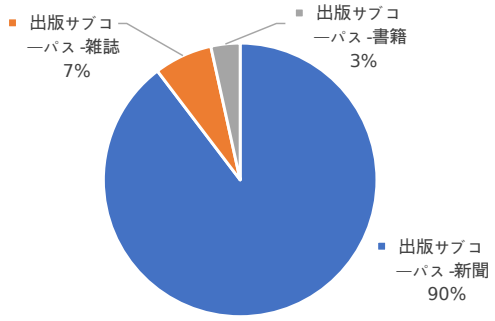
ケーススタディ 2「是非」　図 2b に「是非」という単語の用例に対するクラスタリングの結果を示す。コトバンク辞書と見比べた時、Kmeans によるクラスタリングの結果はクラスタ 0（青）が「ぜひとも」という意味の用例、クラスタ 1（赤）はほとんど「是非を問う」という意味の用例であった。ケーススタディ 1 と同様に、BCCWJ 由来の分野情報と比較するため、図 5a にクラスタ 0（青）「ぜひとも」に含まれる用例の出典分野の割合を示す。これを見ると「ぜひとも」という表現は、口語的な場

sklearn.cluster.KMeans.html

8）https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

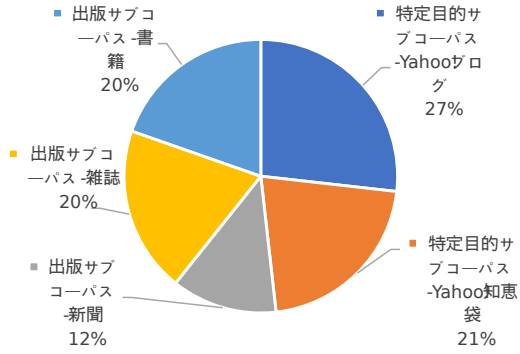


(a) クラスタ 0 に属する用例の各分野の割合

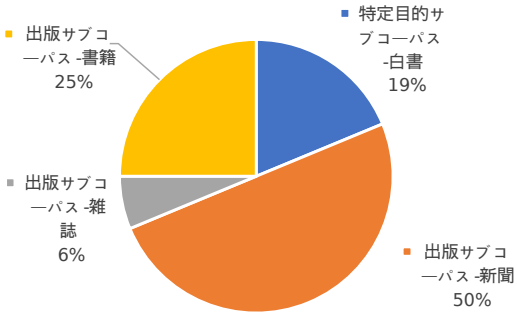


(b) クラスタ 1 に属する用例の各分野の割合

図 4: 「保守」のクラスタリング結果の内訳に対する調査



(a) クラスタ 0 に属する用例の各分野の割合



(b) クラスタ 1 に属する用例の各分野の割合

図 5: 「是非」のクラスタリング結果の内訳に対する調査

面（Yahoo!知恵袋、Yahoo!ブログ）でも書き言葉的な場面（書籍、新聞、雑誌）でもそれぞれ同程度に使用されていることがわかる。また、Yahoo!ブログと Yahoo!知恵袋に属する用例の各意味の使用割合を調べた結果、「是非を問う」という用例は見つけられず、「ぜひとも」が 100%であった。図 5b では、クラスタ 1（赤）「是非を問う」内の用例の出典分野の割合を示す。「是非を問う」は新聞での用例が 50 %を占める一方、25 %が書籍の用例、18.8 %が白書の用例であった。新聞・白書・雑誌での言葉使いは書き言葉の中でも文語体に近く、カジュアルな口語体で書かれる Yahoo!知恵袋や Yahoo!ブログとの明らかな差を観察することができる。またプロットされた各用例を観察したところ、クラスタ 1（赤）「是非を問う」内に「ぜひとも」の用例が 4 つ含まれていたことが分かった。そのため、Kmeans による自動クラスタリングの際に誤分類が起きていることが分かった。ただしここまで述べてきた考察はいずれも直観に反するものでない。そのため、クラスタリング自体に重大な欠陥があるわけではなく、今後の精度向上を期待するものである。

5 おわりに

本研究は分野間で起こる単語の用法の異なりを可視化し、用例研究を支援可能なツールを開発した。このツールを用いることで、コストのかかる統計的な用例分析を行う前に、事前に対象単語が研究目的に適しているかどうか是非の検討が可能になる。今回は Google Colaboratory でデモを公開したが、Google Colaboratory は Python を記述して実行するためのオンライン環境なので、ツールとしては使いにくい。そこで今後の課題として、Web アプリの形式で公開し、クラスタリング以外に分野別の生起率などの情報を可視化しグラフ出力する機能の追加を検討している。用例に対するクラスタリングが適切な分類ができていないことも課題である。NWJC-BERT から得られたベクトルがどの程度クラスタリングに適しているか調査するとともに、クラスタリング手法の改善に取り組んでいきたい。

参考文献

[1] 久屋愛実. 現代書き言葉における外来語の共時的分布 : 「ケース」を事例として. 国立国語研究所論集, Vol. 6, pp. 45–65, 2013.

[2] 呉琳. 〈足を洗う〉という表現が語る言語変化 コーパスによるアプローチ. 言語文化教育研究学会誌『言語文化教育研究』, Vol. 13, pp. 134–168, 2015.

[3] 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, No. 2, pp. 345–371, 2014.

[4] 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.

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

[6] 前川 喜久雄（監修）／伝 康晴・荻野綱男（編）. 講座 日本語コーパス 7 コーパスと辞書. 朝倉書店, 2019.

[7] 保守とは - コトバンク, （2022-1 閲覧）. https://kotobank.jp/word/%E4%BF%9D%E5%AE%88-629994.

A　参考情報

A.1　分野タグ一覧

表 1: BCCWJ に付与されている分野タグ一覧

出版サブコーパス-	書籍 雑誌 新聞
図書館サブコーパス-	書籍
特定目的サブコーパス-	白書 教科書 広報紙 ベストセラー Yahoo!知恵袋 Yahoo!ブログ 韻文 法律 国会会議録

A.2　「出版サブコーパス-新聞」に属する用例の各クラスタの割合

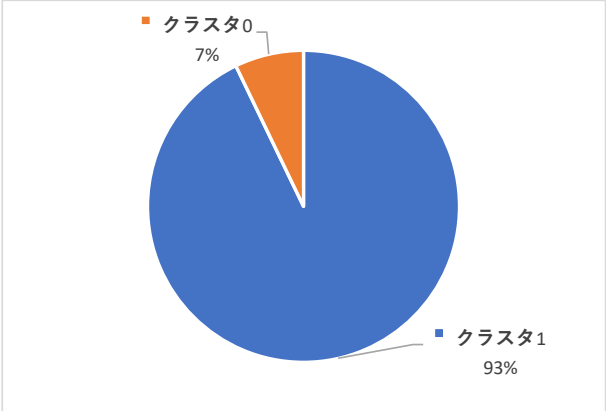


図 6: 「出版サブコーパス-新聞」に属する用例の各クラスタの割合

図 6 では、分野情報が新聞に属する「保守」の用例をまず集め、それらが所属するクラスタの割合を示した。「旧習・伝統を守る」という意味の用例が約 93 %であり、「正常の状態を保つ」の用例が 7 %である。前述の結果と合わせると、「旧習・伝統を守る」という意味の「保守」は分野間で比較しても新聞でよく使用され、新聞内でも「正常の状態を保つ」という意味よりも多く利用されていることが示せた。