日本語意味変化検出の評価セットの拡張と検出手法の評価

凌志棟 [1]　相田太一 [1]　岡照晃 [1]　小町守 [2]
[1] 東京都立大学 [2] 一橋大学
{ling-zhidong,aida-taichi}@ed.tmu.ac.jp
oka-teruaki@tmu.ac.jp　mamoru.komachi@r.hit-u.ac.jp

概要

時代とともに意味が変化する単語をコーパスから自動的に検出・分析する研究は、言語学者だけでなく自然言語処理の研究者からも注目を集めている。英語やドイツ語などの言語では、時期の異なる学習用コーパス（通時コーパス）の公開や評価用単語リストの作成が進んでいるが、日本語では不十分である。本研究では、我々が作成した日本語の評価用単語リストを拡張し、意味変化検出手法の性能評価を行った。頻度に基づく手法をベースラインとして、タイプベースとトークンベースの代表的な手法の性能を比較し、それぞれの手法の特徴を議論した。

1　はじめに

単語の意味は時代とともに変化することがある。従来は言語学者が手作業で検出・分析を行なっていたが、通時コーパスの公開や単語の意味表現の研究の発展により、自動的な検出・分析が可能になったため、近年では自然言語処理の分野でも注目を集めている。これまでに多くの意味変化検出手法が提案されたが、早期の研究では手法の性能評価が論文間で統一されておらず、これらの手法の性能を直接比較できない [1, 2, 3, 4]。また、意味変化の有無に関する情報だけ評価をしているが、意味が変化した単語を全て等しく扱うため、意味変化の程度を考慮した詳細な評価や分析はできない。

この問題に対処するため、Schlechtweg ら [5] は単語の意味変化の度合を計算するフレームワークである Diachronic Usage Relatedness（DURel）を提案した。DURel は、通時コーパスから得られた用例に人手で類似度を付与することで、時間経過に伴う単語の意味変化度合を計算する。現在、さまざまな言語で DURel に基づいた評価用の単語リストが作成・公開されている [6, 7, 8, 9, 10]。

様々な言語で通時コーパスの公開・評価用単語セットの作成が進んでいるが、日本語では通時コーパスの作成が進んでいるものの、評価用の単語セットが十分にない。間淵らが近代から現代にかけて意味が変化した単語のリストを作成した [11] が、リスト内の単語の意味がどの程度変化したのか、という意味変化の度合は付与されていない。本研究では、我々が作成した日本語の評価用単語リスト [12] を拡張し、現在主流となるタイプベース・トークンベースの検出手法について性能評価を行った。その結果、どちらの手法も頻度ベース手法を上回る性能を示した。また、2 つの意味変化の度合によって、その検出の難しさが異なることもわかった。

2　関連研究

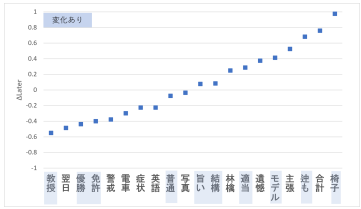
2.1　意味変化検出の手法

意味変化検出では複数時期のコーパスが与えられるが、各時期で独立に単語ベクトルを学習すると異なる空間が得られるため、時期間のベクトルを比較することはできない。そこで、初期化 [1] や対応付け [3] といった様々な手法が提案された。

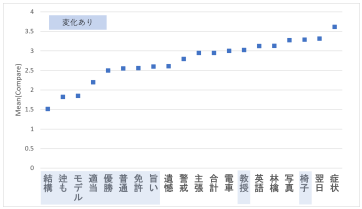
現在は事前学習済みの言語モデルを用いた手法が主流となっており、対象単語の用例ベクトル集合を用いた分析手法が考案された [7]。また、時間情報を考慮したモデルが提案され、非常に高い検出性能が報告されている [13, 14]。

2.2　日本語の意味変化検出

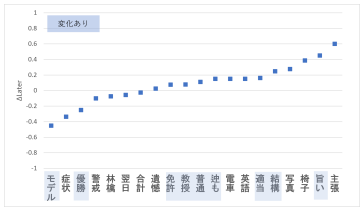
相田ら [15] は、意味変化を測定するための Pointwise Mutual Information (PMI) ベースのモデルを提案し、それらのモデルを英語と日本語のコーパスに適用した。小林ら [16] は、日本語の語義レベルの意味変化を分析するために、辞書ベース [17] とクラスタリングベース [7] の 2 つの異なる手法を適用し、比較した。日本語における先行研究では、意味変化した単語のリスト [11] しかないため、バイナリの評



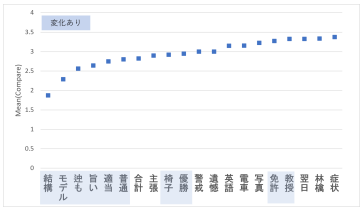
(a) CHJ と BCCWJ の比較で、Δ*Later* の結果。



(b) CHJ と BCCWJ の比較で、Mean(*Compare*) の結果。



(c) SHC と BCCWJ の比較で、Δ*Later* の結果。



(d) SHC と BCCWJ の比較で、Mean(*Compare*) の結果。

図 1: それぞれのコーパスの比較から得られた対象単語の 2 種類の意味変化度合。

価や定性評価に留まっている。

上記の問題に対処するために、我々は日本語の意味変化検出の性能を評価するための評価用単語リスト（Japanese Lexical Semantic Change Detection Dataset; JLSCD）[1] を作成した [12]。この単語リストは明治・大正と平成の 2 つの時期における意味変化を扱ったものである。DURel に基づいており、各単語には意味変化の度合が付与されているため、意味変化検出手法の詳細な性能比較が可能となった。

3 追加データの作成

本研究では、JLSCD に対して**対象単語の拡張**と**比較する時代の追加**を実施した。詳細は A 節に示す。

対象単語 JLSCD では、意味が変化した 6 単語と意味が変化しなかった 3 単語の計 9 単語について、DURel に基づいて意味変化度合を付与した。意味が変化した単語は間淵ら [11] のリストから抽出し、意味が変化しなかった単語はデジタル大辞泉[2] で語義が一つしかない単語をランダムに選定している。本研究では、上記と同様の方法で、意味が変化した単語を 4 単語、意味が変化しなかった単語を 7 単語追加し、合計 20 単語の評価用単語リストを作成した（表 3）。

比較する通時コーパス JLSCD では、日本語歴史コーパス（Corpus of Historical Japanese; CHJ）[3] と現代日本語書き言葉均衡コーパス（The Balanced Corpus of Contemporary Written Japanese; BCCWJ）[4] [18] を用いて、明治・大正と平成の 2 つの時代の比較を行っていた。本研究では、より詳細な時代の比較を行うために、昭和・平成書き言葉コーパス（Showa-Heisei Corpus of written Japanese; SHC）[5] を追加し、昭和と平成の時代も比較した（表 2）。

意味変化度合の算出 DURel フレームワークに基づいて、CHJ と BCCWJ、SHC と BCCWJ で 20 単語に対して 2 種類の意味変化度合を算出した。

- Δ*Later*：各時期のなかで対象単語の用例間の類似度を人手で付与し、時期間でその差分を計算する。絶対値が大きいほど意味の増減が大きいことを示す。
- Mean(*Compare*)：2 つの時期における用例間の類似度を人手で付与し、用例ペアの数で平均を算出する。値が小さいほど意味変化の度合が大きいことを示す。

今回の作業で得られた単語の意味変化度合を図 1 に示す。図より、どちらの時期間でも、Δ*Later* の結果（図 1a、1c）よりも Mean(*Compare*) の結果（図 1b、1d）の方が意味変化した単語を適切に捉えていることがわかる。両者の指標の特徴については、5 節でも考察を行う。

4 検出手法の評価実験

拡張した評価用単語リストを用いて、CHJ と BCCWJ、SHC と BCCWJ のコーパス間で主流の意味変化検出手法の定量評価を行った。今回は前処理として、コーパス内のすべての短単位を語彙素に変

1) https://github.com/tmu-nlp/JapaneseLSCDataset
2) https://japanknowledge.com/en/contents/daijisen/
3) https://clrd.ninjal.ac.jp/chj/
4) https://clrd.ninjal.ac.jp/bccwj/en/index.html
5) https://clrd.ninjal.ac.jp/shc/index.html

換した。また、時期間でデータ量に差があるため、訓練の際に CHJ/SHC をランダムサンプリングして BCCWJ と同等の語数にした。[6)]

4.1 検出手法

ベースライン 頻度差 (Frequency Difference) と共起頻度のベクトル（Count Vector）を使用した。[7)] 頻度差では、対象単語の意味変化度合を両方のコーパス内の出現頻度の差の絶対値で表現する。今回は頻度差の絶対値（*FreqDist*）と対数頻度の差の絶対値（$FreqDist_L$）の 2 つを使用した。Count Vector では、単語ごとに 1 つ Count Vector を学習してから、時期間の対応付けを行って 2 つの時期の対象単語ベクトルの距離を計算する。[8)] 各時期でそれぞれの語彙サイズの次元数を持つ Count Vector に対して、Column Intersection（CI）で時期間の対応付けを取る。Count Vector の軸は対象単語 w の文脈語 c に対応するため、時期間で共通する文脈単語の列を抽出すればよい。対応付けを取った行列に対して、対象単語のベクトルの距離は Cosine Distance（*CosDist*）で計算する。値が大きいほど意味の差が大きい、つまり意味が変化した可能性が高いことを示す。

タイプベースの手法 先行研究で最も性能が良い手法 [19] で日本語データ上の性能を評価する。具体的には、Skip-Gram with Negative Sampling（SGNS）で単語のベクトル表現を学習し、Kim らの Vector Initialization（VI） [1] と Hamilton らの Orthogonal Procrustes（OP）[3] の 2 種類の手法を使用した。[9)] VI は、2 つの時期のデータ上で SGNS を訓練する際に、古い時期のデータで学習済みのモデルで新しい時期のデータを学習するモデルを初期化する。一方、OP は異なる時期データから学習された単語ベクトルを同一空間に対応付けする方法である。時期 a のデータで学習された単語ベクトル行列 $\boldsymbol{W}_a$ を時期 b に対応付けさせる回転行列 $R_{a\to b}$ を以下のように獲得する。

$$\boldsymbol{R}_{a\to b} = \underset{\boldsymbol{R}\boldsymbol{R}^\top=1}{\operatorname{argmin}} ||\boldsymbol{W}_a\boldsymbol{R} - \boldsymbol{W}_b||_F^2 \tag{1}$$

予測では、時期間の単語ベクトルで *CosDist* を計算する。

トークンベースの手法 Giulianelli らが提案した BERT を用いた手法 [7] を評価する。用例から単語ベクトルを獲得するには、浅原らが訓練した NWJC-BERT [20] を使用する。[10)] 対象単語の用例ベクトル集合を獲得してから Kmeans でクラスタリングを行う。K を決める方法として、K を 2 から 10 を試し、最もシルエットスコアが高い K を採用する。2 つの時期のクラスタ分布の差を計算することで単語の意味変化度合を計算する。意味変化度合を計算する方法として、先行研究で高い性能を示した Average pairwise distance（APD）と Jensen-Shannon divergence（JSD）を使用した [7]。

1. **APD**：クラスタリングを用いず、用例ベクトル集合を直接比較する。異なる時期のベクトル集合間の平均距離であり、今回は先行研究で効果のあった Cosine Distance を採用した（$\mathrm{APD}_{CosDist}$） [6]。高い値は意味変化の度合が高いことを意味する。
2. **JSD**：異なる時期のクラスタ分布に依存する。単語 w の時期 a と時期 b の用例クラスタ分布 u_a^w と u_b^w に対して、次のように計算する。

$$\mathrm{JSD}(\mathrm{u}_a^w, \mathrm{u}_b^w) = \mathrm{H}(\frac{1}{2}(\mathrm{u}_a^w + \mathrm{u}_b^w)) - \frac{1}{2}(\mathrm{H}(\mathrm{u}_a^w) - \mathrm{H}(\mathrm{u}_b^w))) \tag{2}$$

H は Boltzmann-Gibbs-Shannon エントロピーである。クラスタ分布が異なっている場合は値が高くなり、意味変化度合が高いことを示す。

4.2 評価

意味変化検出の性能を定量評価するために、拡張した JLSCD を使用する。本研究では、$\Delta Later$ はその絶対値 abs($\Delta Later$) をとり、Mean($Compare$) は負の値 –Mean($Compare$) を取ることで、値の大小と意味変化の大小を対応させた。先行研究 [6] にならい、各手法が予測した意味変化度合と上記の 2 種類の値との Spearman 順位相関係数 ρ で評価する。

5 評価結果と考察

各手法における最もよい性能を出したパラメータの結果を表 1 に示す。CHJ と BCCWJ、SHC と BCCWJ のどちらの比較でも、タイプベース・トークンベースの両方の手法が頻度に基づくベースラインを上回る性能を出した。CHJ と BCCWJ の

6) コーパスをサンプリングしない性能を表 7 に示す。
7) 共通タスク [6] を参考にしている。実装は以下のコードを使用した。https://github.com/Garrafao/LSCDetection
8) 今回は様々なハイパーパラメータで実験を行った（表 6）。
9) ベースラインと同様に、様々なハイパーパラメータで実験を行った（表 6）。
10) 対象単語の用例を BERT に入力して、対象単語トークンの最終層の出力ベクトルを使用する。

表 1: 評価される手法の日本語評価セット上の Spearman 順位相関係数 ρ。太字の数字は各評価値で最もよい結果である。*をつける値は有意確率 $p < 0.05$ である。これは、「意味変化度合のランキングに対して、人手評価と検出手法の予測は無相関である」が生じる確率 $p < 0.05$ であり、棄却できることを意味する。

モデル	時期対応付け	距離計算	CHJ と BCCWJ(ρ)		SHC と BCCWJ(ρ)	
			abs($\Delta Later$)	−Mean($Compare$)	abs($\Delta Later$)	−Mean($Compare$)
正規化頻度		$FreqDist$	0.018	0.441	0.114	0.034
log 正規化頻度		$FreqDist_L$	-0.047	0.273	0.107	-0.227
Count Vector	CI	$CosDist$	0.317	0.310	0.220	-0.384
SGNS	OP	$CosDist$	**0.553***	0.742*	0.232	0.392
SGNS	VI	$CosDist$	0.481*	0.744*	**0.339**	**0.632***
BERT	-	$APD_{CosDist}$	0.276	**0.748***	0.035	0.572*
BERT	-	Kmeans+JSD	0.098	0.308	0.017	0.363

比較では、abs($\Delta Later$) ではタイプベースの手法、−Mean($Compare$) ではトークンベースの手法が最もよい性能を出した。一方で、SHC と BCCWJ の比較では、両方の意味変化度合に対してタイプベースの手法が最もよい性能を出した。小林ら [16] が現代語 BERT の Fine-Tuning により検出性能が向上したが、SemEval2020 Task 1 [6] より、Fine-Tuning したトークンベースのモデルはタイプベースのモデルを上回る可能性は低い。他の手法として、時間を考慮したトークン [13] や注意機構 [14] を導入することでタイプベースを上回ることが報告されている。また、多言語の XLM-R モデルを用いた XL-LEXEME [21] は英語やドイツ語で最高性能を達成した。これらの手法を日本語で評価することは、今後の課題である。

CHJ と BCCWJ の比較での性能は、SHC と BCCWJ での比較の性能を上回る傾向を示している。原因として、表 2 より、CHJ と BCCWJ の時期の差が SHC と BCCWJ より長く、対象単語の意味や文脈が顕著に異なるため、意味変化を捉えやすくなると考える。

ここで、2 つの意味変化度合の結果を比較すると、abs($\Delta Later$) での結果はどちらの比較でも統計的有意の結果が少ないことがわかる。この結果は、Kutuzov ら [8] の結果と一致する。原因として、−Mean($Compare$) は 2 つの時期で意味の違いを直接比較するのに対し、abs($\Delta Later$) では時期間の意味の増減を相対的に扱っていることが挙げられる。今回実験で使用した手法はどれも時期間の意味の増減を扱っておらず、2 つの時期における意味の違いを直接比較するため、−Mean($Compare$) での性能が abs($\Delta Later$) での性能を上回ったのだと考える。

タイプベースのモデルを用いた手法では、SGNS+VI がほとんどの評価で SGNS+OP を上回った。Schlechtweg ら [19] によるドイツ語での結果では、SGNS+OP の組み合わせが他の手法を上回る性能を出したが、今回は同じ結果は得られなかった。この要因として、対応付けの前提条件が挙げられる。ベクトル空間の対応付け手法である OP を適用するには、学習文書データ内のほとんどの単語が意味変化しないことが前提となる。今回の学習データが対応する明治から平成までの日本語は、単語の語義以外に、表記や文体も大きく変化したため、OP の前提条件を満たせない可能性が高いと考える。この傾向は、相田ら [22] の実験結果からも見られる。

トークンベースのモデルを用いた手法は、$APD_{CosDist}$ を用いた距離計算が JSD をはるかに上回った。日本語以外の多くの言語のデータセット上でも、APD が JSD を上回る性能を出すケースが多いことを Montanelli ら [23] により報告されている。JSD などの手法はクラスタリングの性能にも依存するため、単純な類似度計算より対象単語の意味変化を検出しにくい傾向があると考える。

6 おわりに

本研究は意味変化検出の評価セットを拡張した上で、主流の意味変化検出手法が日本語における性能評価を行った。その結果、異なる意味変化度合の測定値に対して、タイプベースのモデルを用いた手法とトークンベースのモデルを用いた手法は異なる性能を出した。Mean($Compare$) では人間の判断との有意な相関を達成したが、$\Delta Later$ で性能が相対的に低いことを示した。今後は時間を考慮した注意機構や XL-LEXEME の日本語における性能評価を行いたい。また、今回の実験結果は、日本語あるいはすべての言語の通時的な意味変化に興味を持つ研究者の参考になることを期待している。

謝辞

本研究は、国立国語研究所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」および JST さきがけ「文理融合による人と社会の変革基盤技術の共創」JPMJPR2366 の成果の一部を含むものである。

参考文献

[1] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In Cristian Danescu-Niculescu-Mizil, Jacob Eisenstein, Kathleen McKeown, and Noah A. Smith, editors, **ACL 2014 Workshop on Language Technologies and Computational Social Science**, pp. 61–65, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

[2] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In **The 24th WWW**, WWW '15, p. 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[3] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In **ACL**, pp. 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Lea Frermann and Mirella Lapata. A Bayesian model of diachronic meaning change. **TACL**, Vol. 4, pp. 31–45, 2016.

[5] Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In **NAACL**, pp. 169–174, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[6] Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In **The 14th SemEval**, pp. 1–23, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[7] Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In **ACL**, pp. 3960–3973, Online, July 2020. Association for Computational Linguistics.

[8] Julia Rodina and Andrey Kutuzov. RuSemShift: a dataset of historical lexical semantic change in Russian. In **COLING**, pp. 1037–1047, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[9] Andrey Kutuzov and Lidia Pivovarova. Three-part diachronic semantic change dataset for Russian. In **LChange**, pp. 7–13, Online, August 2021. Association for Computational Linguistics.

[10] Jing Chen, Emmanuele Chersoni, and Chu-ren Huang. Lexicon of changes: Towards the evaluation of diachronic semantic shift in Chinese. In **LChange**, pp. 113–118, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[11] 間淵洋子, 小木曽智信. 近現代日本語の意味変化分析のための単語データセット構築の試み. 言語処理学会第 27 回年次大会発表論文集, pp. 1166–1170, 2021.

[12] Zhidong Ling, Taichi Aida, Teruaki Oka, and Mamoru Komachi. Construction of evaluation dataset for japanese lexical semantic change detection. In **PACLIC**, Hong Kong, December 2023. Association for Computational Linguistics.

[13] Guy D. Rosin, Ido Guy, and Kira Radinsky. Time masking for temporal language models. In **WSDM**, pp. 833–841, New York, NY, USA, 2022. Association for Computing Machinery.

[14] Guy D. Rosin and Kira Radinsky. Temporal attention for language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Findings of NAACL 2022**, pp. 1498–1508, Seattle, United States, July 2022. Association for Computational Linguistics.

[15] 相田太一, 小町守, 小木曽智信, 高村大也, 持橋大地. 異なる時期での意味の違いを捉える単語分散表現の結合学習. 自然言語処理, Vol. 30, No. 2, pp. 275–303, 2023.

[16] 小林千真, 相田太一, 岡照晃, 小町守. Bert を用いた日本語の意味変化の分析. 自然言語処理, Vol. 30, No. 2, pp. 713–747, 2023.

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

[18] 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.

[19] Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In **ACL**, pp. 732–746, Florence, Italy, July 2019. Association for Computational Linguistics.

[20] 浅原正幸, 西内沙恵, 加藤祥. Nwjc-bert: 多義語に対するヒトと文脈化単語埋め込みの類似性判断の対照分析. 言語処理学会第 26 回年次大会発表論文集, pp. 961–964, 2020.

[21] Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **ACL**, pp. 1577–1585, Toronto, Canada, July 2023. Association for Computational Linguistics.

[22] 相田太一, 小町守, 小木曽智信, 高村大也, 持橋大地. 異なる時期での意味の違いを捉える単語分散表現の結合学習. 自然言語処理, Vol. 30, No. 2, pp. 275–303, 2023.

[23] Stefano Montanelli and Francesco Periti. A survey on contextualised semantic shift detection, 2023.

A　DURel に基づくデータの構築

対象コーパスの概要を表 2 に、対象単語の一覧を表 3 に示す。単語用例の収集は CHJ,SHC,BCCWJ の 3 つのコーパスを中納言オンラインコーパス検索エンジンでサンプリングした[11]。サンプリングする際に、異なるコーパスの比較に応じて 3 つのグループ *Eariler*、*Later*、*Compare* に分ける。*Earlier* のペアには CHJ/SHC の用例が 2 つ、*Later* のペアには BCCWJ の用例が 2 つ、*Compare* のペアには CHJ/SHC と BCCWJ の用例がそれぞれ 1 つ含まれる。その後、各グループの用例ペアに対して、表 4 のスコアリングで意味的類似度をアノテーションした。アノテータは日本語に関する豊富な知識を持つ国立国語研究所の研究者である[12]。各グループのアノテーション結果に基づいて、各対象単語 w の意味変化の度合を計算する。

- $\Delta Later_w$ = Mean($Later_w$) − Mean($Earlier_w$): *Earlier* のアノテーションスコアの平均から、*Later* のアノテーションスコアの平均を引いたものである。この測定値は *Earlier* グループから *Later* グループへの意味の関連性の変化を測定する。正の値は語義が増加し、対象単語が語義増加を示し、負の値は語義減少を示す。
- Mean($Compare_w$): *Compare* グループのアノテーションスコアの平均を計算する。スコアが高い/低いほど、変化の種類に関係なく、2 つの時期間の変化が弱い/強いことを示す。

今回拡張した評価セットの統計量を表 5 に示す。

表 2: コーパスの概要。今回は全て雑誌を使用した。

コーパス	時期	トークン数	タイプ数
CHJ	1874–1925	11.0M	96K
SHC	1925–2000	27.3M	116k
BCCWJ	2001–2005	5.2M	61K

表 3: 拡張された評価セットの対象単語、太字は新しく追加した単語である。

変化あり	結構	**モデル**	**迎も**	**旨い**	適当
	普通	椅子	優勝	免許	教授
変化なし	**症状**	主張	**警戒**	林檎	**合計**
	翌日	**遺憾**	**電車**	**英語**	写真

11）https://chunagon.ninjal.ac.jp/
12）https://www.ninjal.ac.jp/

表 4: アノテーションで使用する意味類似度に基づくスコアリング。

スコア	類似度
4	意味が一致した
3	意味が近い
2	意味が少し関連している
1	対象単語の意味が完全に異なる
N/A	判断できない

表 5: 拡張された評価セットの概要。$|V|$ は単語数、n はアノテータの数、$|Pair|$ は用例ペアの数、$|Score|$ はアノテーションされた類似度の総数を示す。α と ρ はそれぞれ一致率の Krippendorff's Alpha と Spearman 順位相関係数である。

比較コーパス	$\|V\|$	n	$\|Pair\|$	$\|Score\|$	α	ρ
CHJ, BCCWJ	20	4	1200	3480	0.280	0.328
SHC, BCCWJ	20	2	1200	2400	0.258	0.302

B　実験設定

ベースラインとタイプベースの学習パラメータを表 6 に示す。また、サンプリングをせずにコーパス間のデータ量を揃えないときの結果を表 7 に示す。

表 6: Count Vector と Skip-gram の学習パラメータ。

モデル	次元数	窓幅	エポック数
Count Vector	50,100,200,	5,10,15,	-
SGNS	300,350,400	20,30	30

表 7: 全てのデータを使った時の Spearman 順位相関係数 ρ。*をつける値は $p < 0.05$ を示す。

(a) CHJ と BCCWJ

モデル	abs($\Delta Later$)	−Mean($Compare$)
FreqDist	-0.102	0.258
$FreqDist_L$	-0.141	0.066
CV+CI+*CosDist*	0.354	0.375
SGNS+OP+*CosDist*	0.574*	0.766
SGNS+VI+*CosDist*	0.583*	0.756

(b) SHC と BCCWJ

モデル	abs($\Delta Later$)	−Mean($Compare$)
FreqDist	-0.146	0.145
$FreqDist_L$	-0.228	0.125
CV+CI+*CosDist*	-0.030	-0.376
SGNS+OP+*CosDist*	-0.032	0.325
SGNS+VI+*CosDist*	0.031	0.529