nwjc2vec:『国語研日本語ウェブコーパス』 に基づく 単語の分散表現データ

†浅原 正幸 (人間文化研究機構 国立国語研究所)
岡 照晃 (人間文化研究機構 国立国語研究所)

†masayu-a@ninjal.ac.jp

1 はじめに

本発表では我々が構築した『国語研日本語ウェブコーパス』(以下，NWJC)[2] に基づく単語の分散表現データ nwjc2vec について紹介する。具体的には NWJC (2014 年第 4 四半期データ) 258 億語から、word2vec [3] のモデルを表層形のみのものと UniDic 形態論情報つきのものの 2 種類作成した。

以下、2 節では NWJC の概要と word2vec に与えた各種パラメータについて示す。3 節では UniDic-分類語彙表番号対応表に基づく評価を示す。最後にまとめと今後の計画について示す。

2 NWJC からの分散表現の構築

2.1 NWJC の概要

NWJC はウェブを母集団とし、100 億語規模を目標として構築した日本語コーパスである。ウェブアーカイブの構築で用いられる Heritrix-3.1.1 [1] クローラを運用することで、1 年間、3 か月おきに、固定した約 1 億 URL のウェブページを収集している。得られたウェブページは nwc-toolkit-0.0.2 [2] を用いて、日本語文抽出と正規化を行う。コピーサイトの問題を緩和するために、文単位の単一化（文の異なりを用いる）を行った。アノテーションはすべて自動解析を用い、形態論情報、および係り受け情報を付与している。形態素解析には形態素解析器 MeCab-0.996[3] と UniDic-2.1.2[4] を使用し、係り受け解析には係り受け解析器 CaboCha-0.69[5] と UniDic 主辞規則 [6] を使用した。

収集したデータを研究者に提供することが求められているが、著作権の問題があり、収集したデータをそのまま外部の研究者に提供することは難しい。そこで、文字列のみならず、形態論情報や係り受け構造に基づく検索系を構築し、例文とともに元データが含まれる URL へのリンクを含めて提示するサービスを構築した [1]。

このサービスから利用可能なデータは、2014 年 10-12 月期収集データ (NWJC-2014-4Q データ) に基づく。基礎統計は表 1 のとおりである。

表 1: 基礎統計:NWJC-2014-4Q データ

収集 URL 数	83,992,556	8399 万 URL
文数（のべ数）	3,885,889,575	38 億文
文数（異なり数）	1,463,142,939	14 億文
国語研短単位数	25,836,947,421	258 億単位

2.2 訓練のパラメータ

表 1 に示した NWJC-2014-4Q データを用いて分散表現データを構築する。分散表現データの構築には word2vec [7] の CBOW モデルを用いた。表 2 に word2vec 実行時の各種パラメータを示す。

学習時のトークンには、書字形出現形のみを使った word と、形態論情報 [8] を含めた mrph の 2 種類のモデルを用意した。

3 構築した分散表現の評価

3.1 評価方法

評価は形態論情報付きデータ mrph について行う。NWJC 中に出現した UniDic 体系の品詞「形容詞」

[1] http://webarchive.jira.com/wiki/display/Heritrix/Heritrix/
[2] 現在公開停止。
[3] https://taku910.github.io/mecab/
[4] https://osdn.jp/projects/unidic/
[5] https://taku910.github.io/cabocha/
[6] ./configure --with-posset=UNIDIC
[7] https://github.com/svn2github/word2vec
[8] unidic-mecab_kana-accent-2.1.2 の dicrc に記載の素性 26 種。

表 2: word2vec の実行時のパラメータ

CBOW or skip-gram	-cbow	1
次元数	-size	200
文脈長	-window	8
負サンプリング数	-negative	25
階層化 softmax	-hs	0
最低頻度閾値	-sample	1e-4
反復回数	-iter	15

4,188 件（UniDic 既登録語）を対象語とし、対象語と分散表現の類似度が高い上位 40 語（抽出語）を抽出する。

抽出語が UniDic に登録されている場合、『形態素解析辞書 UniDic と分類語彙表の見出し語対応付けデータ』（以下「対応表」）[5] を用いて、可能な分類語彙表番号を枚挙する。対応表に UniDic の語彙素番号 [4] が登録されていない場合は分類語彙表番号が割り当てられない。

評価は、単語対（対象語と抽出語）に UniDic 語彙素番号と少なくとも 1 つ以上の分類語彙表番号が付与されている場合に実施する。UniDic の語彙素番号を共有しているものを最も近いとし、その後分類語彙表番号のうちの分類番号の桁数の一致度を参照していく。

尚、今回評価対象にしなかった UniDic 未登録語および分類語彙表番号なしの語については、今後、本データの類似度に基づいて、人手で登録する作業を実施する予定である。

分類語彙表番号は以下のような構造をしている。

例：「昨年」(分類番号：1.1642)

類	部門	中項目	分類項目
体 (1)	関係 (.1)	時間 (.16)	過去 (.1642)

分類番号	段落番号	小段落番号	語番号
1.1642	01	01	01

この分類番号が評価対象になり、その桁数に応じて、類 (1 桁)・部門 (2 桁)・中項目 (3 桁)・分類項目 (5 桁) と概念が細かくなる。正確には、類は統語的な性質（体・用・相・その他）を表し、ピリオド以下の部門・中項目・分類項目が意味を表す。しかしながら、本稿では簡単に分類語彙表の分類番号の上位からの一致桁数で評価した。段落番号・小段落番号・語番号は『分類語彙表』の紙面上の配置の情報なので、評価に利用しない。

単語対のいずれかに 2 つ以上分類語彙表番号が付与されている場合には、全ての組み合わせで最もよい結果をその単語対の UniDic/分類語彙表番号により定義されるの意味的類似度により評価する。

以下に例を示しながら、評価手順について詳説する。

- 評価対象外
  - 語彙素番号なし
    UniDic に登録されてない語
  - 分類語彙表番号なし
    UniDic に登録されているが、UniDic 語彙素番号-分類語彙表番号対応表にない語
    * 語彙素不一致
      語彙素番号が異なる対
      「なれなれしく」⇔「よそよそしく」
      (類似度の値:0.699, 類似度の順位:7 位)
    * 語彙素一致
      語彙素番号が同じ対
      「濃ゆく (語彙素番号 247329, 分類語彙表番号 未定義)」
      ⇔「濃いく (同 247329, 分類語彙表番号 未定義)」
      (0.862, 1 位)
- 評価対象
  - 語彙素一致
    語彙素番号が同じ対
    「スゲー (語彙素番号 19163, 分類語彙表番号 3.1400:3.3012:3.3030)」
    ⇔「すんげー (19163, 3.1400:3.3012:3.3030)」
    (0.883, 4 位)
  - 分類項目一致 (類を含む)
    語彙素番号が異なるが、分類語彙表番号上位 5 桁まで一致する対
    「苛立たしい (47395, 3.3013)」
    ⇔「はがゆい (29538, 3.3013)」
    (0.596, 15 位)
  - 中項目一致 (類を含む)
    語彙素番号が異なるが、分類語彙表番号上位 3 桁目まで一致する対
    「ニクい (28233, 3.3020)」
    ⇔「シブい (15685, 3.3011:3.3300:3.5030:3.5050)」
    (0.596, 35 位)
  - 部門一致 (類を含む)
    語彙素番号が異なるが、分類語彙表番号上位 2 桁目まで一致する対
    「ユルく (38933, 3.1341:3.1800)」
    ⇔「ヌルく (28745, 3.1915:3.3680:3.5170)」
    (0.595, 10 位)
  - 類一致
    語彙素番号が異なるが、分類語彙表番号上位 1 桁目まで一致する対

「うっとおしく (3266, 3.3014:3.5150)」
⇔「薄ら寒く (47221, 3.1915)」
(0.593, 33 位)

– 不一致
語彙素番号も分類語彙表も異なる対
「うらやましー (3472, 3.1302:3.3020)」
⇔「わあー (41394, 4.3000:4.3010)」
(0.593, 38 位)

評価は、「類似度の値に基づく評価」と「類似度の順位に基づく評価」の 2 種類実施する。類似度の値に基づく評価においては、類似度を 0.1 刻みの Bucket 毎に、上記評価の分布をみる。類似度の順位に基づく評価においては、各形態素に対する類似度順毎に、上記評価の分布をみる。

3.2 類似度に基づく評価の結果

表 3 に類似度の値に基づく評価を示す。類似度 0.9 以上のもののほとんどが語彙素が一致するレベルであった。

【語彙素一致】
「きつい, 形容詞, 一般,*,*, 形容詞, 連体形-一般, キツイ, きつい (8687, 3.1341:3.1400:3.1912)」
⇔「キツイ , 形容詞, 一般,*,*, 形容詞, 連体形-一般, キツイ, きつい (8687, 3.1341:3.1400:3.1912)」
(0.967, 1 位)

【語彙素一致】
「やばい, 形容詞, 一般,*,*, 形容詞, 連体形-一般, ヤバイ, やばい (38389, 3.1346)」
⇔「ヤバい , 形容詞, 一般,*,*, 形容詞, 連体形-一般, ヤバイ, やばい (38389, 3.1346)」
(0.912, 2 位)

類似度 0.9 以上の不一致の事例は形態素解析の誤解析由来のものであったが、統語的な情報を表す「類」以外の情報（「部門」・「中項目」・「分類項目」）は一致していた。

【不一致】
「黒けれ, 形容詞, 一般,*,*, 形容詞, 仮定形-一般, クロイ, 黒い」(10575, 3.5020:3.5060)
⇔「白けれ , 動詞, 一般,*,*, 下一段-カ行, 仮定形-一般, シラケル, 白ける」(17106,2.3011:2.5020)
(0.914, 1 位)

類似度 0.8 以上の事例は言い換え可能なレベルの同義表現が確認できた。

【分類項目一致】
「愛くるしい, 形容詞, 一般,*,*, 形容詞, 連体形-一般, アイクルシイ, 愛くるしい (46360, 3.3020)」
⇔「愛らしい , 形容詞, 一般,*,*, 形容詞, 連体形-一般, アイラシイ, 愛らしい (230, 3.3020)」
(0.887, 8 位)

【部門一致】
「辛う, 形容詞, 一般,*,*, 形容詞, 連用形-ウ音便, ツライ, 辛い (4800, 3.3014)」
⇔「厳しゅう , 形容詞, 一般,*,*, 形容詞, 連用形-ウ音便, キビシイ, 厳しい (8793, 3.1346:3.3680:3.5150)」
(0.808, 3 位)

表 4 に類似度の順位に基づく評価を示す。類似度の順位が 1 位のものには以下のような例が確認された。

【分類項目一致】
「厚かまし, 形容詞, 一般,*,*, 形容詞, 連用形-一般, アツカマシイ, 厚かましい (814, 3.3041)」
⇔「図々しく , 形容詞, 一般,*,*, 形容詞, 連用形-一般, ズウズウシイ, 図々しい (19611, 3.3041)」
(0.909, 1 位)

以下の事例では、形態素解析の誤りではあるが、対義語相当の表現が得られている。

【類一致】
「弱き, 形容詞, 一般,*,*, 文語形容詞-ク, 連体形-一般, ヨワイ, 弱い (39541, 3.1400:3.5710)」
⇔「強き , 名詞, 普通名詞, 形状詞可能,*,*,*, ツヨキ, 強気 (24793, 1.3000:1.3420:3.3000:3.3420)」
(1 位, 0.856)

文語的な表現も含まれている一方、これらについては Web 上に事例が少ないためか、文体的な類似度が反映されているのではないかと考える。

【不一致】
「早き, 形容詞, 一般,*,*, 文語形容詞-ク, 連体形-一般, ハヤイ, 早い (30135, 3.1660:3.1913)」
⇔「落つる, 動詞, 一般,*,*, 文語上二段-タ行, 連体形-一般, オチル, 落ちる (5028, 2.1251:2.1540:2.1584:2.1931:2.3321:2.5701)」
(0.734, 1 位)

【不一致】
「無う, 形容詞, 非自立可能,*,*, 形容詞, 連用形-ウ音便, ナイ, 無い (27442, 3.1200)」
⇔「僻事 , 名詞, 普通名詞, 一般,*,*,*, ヒガゴト, 僻事 (64449, 1.3046)」
(0.659, 1 位)

表 3: 類似度の値に基づく評価 (「分類語彙表番号」ありが評価対象)

類似度	語彙素番号なし	分類語彙表番号なし		分類語彙表番号あり						
		語彙素 不一致	語彙素 一致	語彙素 一致	分類項目 一致	中項目 一致	部門 一致	類 一致	不一致	合計
0.9	71	33	44	384	57	5	6	0	1	454
0.8	327	719	69	741	698	399	569	509	133	3232
0.7	1885	3680	188	1647	1812	1752	2623	2622	2759	14129
0.6	5673	10091	186	2262	2933	2017	3223	3066	7094	21728
0.5	14163	15058	117	1987	2934	1743	3454	3095	10862	25126
0.4	19822	10792	32	718	957	718	1532	1305	7575	13318
0.3	3634	1473	1	42	58	56	137	122	1013	1456
0.2	9	1	0	0	0	0	1	1	7	9
計	45584	41847	637	7781	9449	6690	11545	10720	29444	79452

表 4: 類似度の順位に基づく評価 (「分類語彙表番号」ありが評価対象)

順位	語彙素番号なし	分類語彙表番号なし		分類語彙表番号あり						
		語彙素 不一致	語彙素 一致	語彙素 一致	分類項目 一致	中項目 一致	部門 一致	類 一致	不一致	合計
1	1112	794	77	949	329	83	174	158	464	2205
2	1156	827	66	693	330	129	208	177	530	2139
3	1078	914	43	555	359	141	250	206	568	2153
4	1076	984	36	448	336	166	231	225	592	2092
5	1142	987	30	404	285	176	233	234	596	2029
6	1086	1007	26	338	278	165	286	277	634	2069
7	1096	1028	29	295	286	163	279	257	662	2035
8	1114	1011	19	287	272	159	265	270	690	2044
9	1107	1043	27	247	246	173	284	273	694	2011
10	1096	1055	12	238	271	190	272	269	673	2025

4 おわりに

本稿では『国語研日本語ウェブコーパス』に基づく単語の分散表現データを紹介した。形容詞 (相の類) を中心に、分類語彙表番号に基づく定量的な評価を行った。CBOW モデルをお求めの方は第一著者に連絡されたい。

今後、Skip-gram モデルについても、構築次第公開する予定である。

謝辞

本研究の一部は国語研コーパス開発センター「超大規模コーパス」プロジェクト (2011-2015)・コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021)・言語変化研究領域共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」(2016-2021) によるものです。

参考文献

[1] Masayuki Asahara, Kazuya Kawahara, Yuya Takei, Hideto Masuoka, Yasuko Ohba, Yuki Torii, Toru Morii, Yuki Tanaka, Kikuo Maekawa, Sachi Kato, and Hikari Konishi. 'bonten' – corpus concordance system for 'ninjal web japanese corpus'. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pp. 25–29, 2016.

[2] Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. Archiving and analysing techniques of the ultra-large-scale web-based corpus project of ninjal, japan. Alexandria: The Journal of National and International Library and Information Issues, Vol. 25, No. 1-2, pp. 129–148, 2014.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In ICLR Workshop paper, 2013.

[4] 小木曽智信, 中村壮範. 『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用. 自然言語処理, Vol. 21, No. 2, pp. 301–332, 2014.

[5] 近藤明日子, 田中牧郎. 分類語彙表・unidic 見出し対応表の構築—コーパスへの網羅的・系統的な語義情報付与を目指して—. 言語処理学会第 23 回年次大会発表論文集, 2017.