分散表現：単語分割の変化による影響調査

勝田 哲弘　山本 和英
長岡技術科学大学
{katsuta, yamamoto}@jnlp.org

1　はじめに

単語の分散表現は自然言語処理の多くのタスクで使用されており、その評価が重要視されてきている。単語の分散表現の簡易的な評価として人間が作成した単語類似度データセットとの相関を調べる方法が一般的に用いられている [1]。類似語データセットは英語では、SimLex-999[2], WS353[3], RG[4], RW[5], VSD[6] 等、多くのデータセットが作られている。近年では、類似性と関連性を明確に分けることの必要性が指摘されており、SimLex-999 では明確に類似性と関連性を区別したデータセットの作成を行っている。区別されているデータセットでは NER や品詞付与の精度との間に正の相関が得られている [7]。

日本では Sakaizawa らが公開している日本語単語類似度データセット（JWSD）[8] や猪原らによって日本語類似度･関連度データセット（JWSAN）[9] が作成されている。JWSD では日本語語彙平易化の評価用データセットから単語の対を抽出し、高頻度のものだけでなく低頻度のものにも注目している。JWSAN では WordNet[1] の Synset や PPMI 類似度を基に類似度と関連度の高低が広く分布するデータセットの作成をしている。また、単語分散表現の評価として推奨している JWSAN-1400 では類似度、関連度の分布の偏りを少なくしたデータセットになっている。それぞれのデータセットの規模は表 1 に示す通りである。

この研究では単語分散表現を学習する上で効果的な単語の定義について調査を行った。UniDic 辞書[2]を用いた MeCab[10] による形態素解析結果を基により学習に有利になる単語の集合を調査する。単語分割の単位は品詞の内容語と機能語に注目し、内容語に機能語を結合することで分割の単位を変化させる。類似度データセットとの相関が高い分散表現を作ることを目的とし、分散表現の改善を行うことで実際の NLP タスクの精度向上に繋げたいと考えている。

表 1: データセットの品詞ごとのペア数

データセット	名詞	動詞	形容詞	副詞
JWSD	1103	1464	960	902
JWSAN	1096	232	72	-

2　関連研究

一般的に単語分散表現の学習は word2vec[11] や Glove[12]、fastText[13] で用いられているように skip-gram や continuous bag-of-words (cbow) モデルによって学習される。fastText では、単語表現は文字 n-gram の合計で計算される。サブワード情報を利用することが単語間の n-gram の共有を可能にし、効率的に訓練されることを示している。

吉井ら [14] は単語類推タスクと文完成タスクの 2 種類で単語の分散表現の評価を行っている。単語類推タスクでは、表記ゆれや単語の活用形が与える影響について、日本語単語類似度データセットで評価を行う。単語の分散表現は、公開されている日本語の訓練済み単語分散表現に加え、word2vec と Glove を用いて実験を行う。評価は, 人手でアノテーションされた単語の類似度と単語間の分散表現のコサイン類似度を、スピアマンの順位相関係数によって行う。英語の単語類推タスクよりも正答率が低くなっている。

単語の分割による分散表現の影響としては形態素解析器を用いない文字 n-gram での分散表現を構築する研究があり、頻度や期待単語頻度を用いた単語の分割方法が分散表現の学習に大きく影響していることが分かる。[15][16]

[1] http://compling.hss.ntu.edu.sg/wnja/
[2] https://unidic.ninjal.ac.jp/

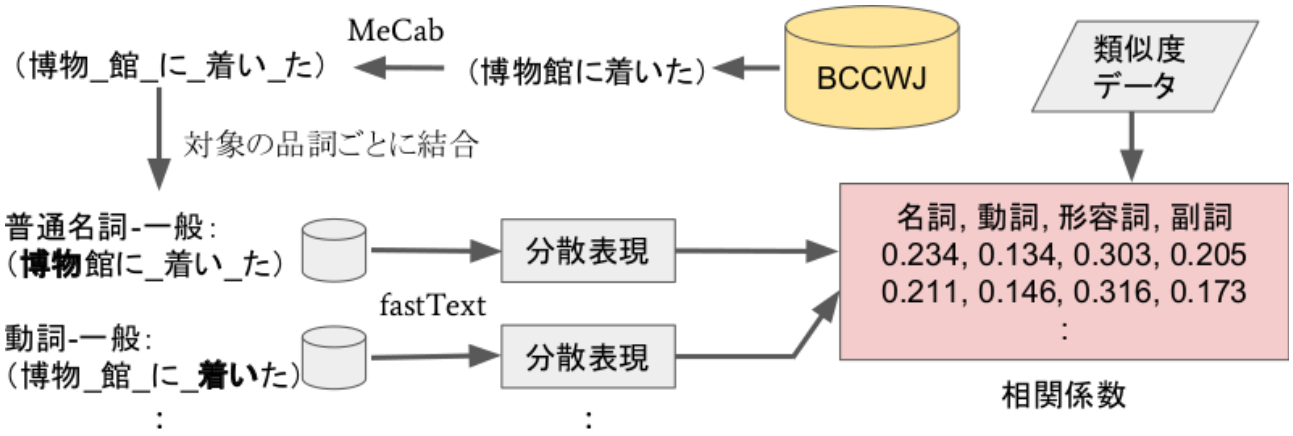

図 1: 実験の流れ

表 2: 形態素解析のみを行った場合の評価

品詞	名詞	動詞	形容詞	副詞
相関係数	0.235	0.160	0.303	0.186

3 方法

3.1 単語分散表現の学習

学習に使用するデータは現代日本語書き言葉均衡コーパス (BCCWJ) からランダムに 500,000 文を抽出したものを使用する。

学習には fastText をデフォルト設定のまま使用する。単語ベクトルは学習したモデルから予測されるものを使用する。

3.2 分散表現の評価方法

データの規模は表 1 に示すように JWSD のほうがデータ数が多いため、本実験では評価に JWSD を使用する。

類似語データセットでは 2 つの単語間の類似度が人手によって付与されている。分散表現の評価では、テキストから分散表現を学習したモデルが予測する単語ベクトルから単語間のコサイン類似度を計算し、人手で付与された類似度との相関を調べる。相関が高いほど、人の感覚に近い分散表現が生成できていることになり、より良い単語ベクトル空間を形成できていると考えられる。

相関の計算では、スピアマンの順位相関係数 ρ を計算する。スピアマンの相関は、2 つの連続変数または順位変数間の単調関係を評価する。単調関係では変数が一緒に変化するが、その値が一定の割合とは限らない。スピアマンの相関係数は、生データではなく各変数の順位値に基づき評価される。

3.3 単語単位の変更

日本語では、サ変名詞のように後続する単語によって使われ方が異なる品詞がある（「サ変名詞＋する」のように）。機能語を結合する目的はそのような単語を機能語の結合で明確に分けることであり、サ変名詞、サ変動詞等の語義曖昧性を解消することで、分散表現の改善につながるのでないかと考えている。

方法としては図 1 に示すように、MeCab によって UniDic 単位で区切られた単語に対して内容語と機能語で品詞を区別し、内容語に続く機能語をひとまとめにすることで単語の分割単位を変更させ、そのときの分散表現の変化を調べる。変更を行う内容語は 1 つの品詞ごとであり、動詞一般を対象とするときは、名詞や副詞には何も変更を行わない。

内容語と機能語は以下のように定義し、内容語の品詞は UniDic の小分類を使用し、機能語は大分類で品詞を分類している。

- 内容語：名詞、代名詞、形状詞、副詞、感動詞、動詞、形容詞
- 機能語：非自立、連体詞、接頭辞、接尾辞、助動詞、（助詞、接続詞）

機能語は基本的に内容語以外の品詞全てを対象にしているが、必要以上の結合は不当に学習を困難にするため、助詞と接続詞を機能語として定義した場合（1）と機能語から除外した場合（2）で実験を行うことにした。本実験では、MeCab で解

表 3: 各品詞ごとに単語を加工した際のスピアマンの順位相関係数

品詞			名詞		動詞		形容詞		副詞	
大分類	中分類	小分類	機能語 1	機能語 2	機能語 1	機能語 2	機能語 1	機能語 2	機能語 1	機能語 2
代名詞			*0.230	0.235	*0.143	0.161	*0.316	*0.315	*0.210	0.184
副詞			0.2334	*0.243	*0.133	*0.142	*0.319	*0.319	*0.240	*0.196
動詞	一般		0.235	*0.238	*0.152	*0.163	*0.310	*0.319	*0.171	*0.175
名詞	助動詞語幹		*0.238	0.235	0.159	*0.158	*0.314	*0.316	*0.182	0.184
	数詞		0.233	*0.227	*0.147	*0.159	*0.309	*0.315	*0.200	*0.194
	普通名詞	サ変	*0.227	*0.225	*0.140	*0.164	*0.305	*0.317	*0.191	*0.192
		サ変形状詞	0.234	*0.238	*0.168	*0.165	*0.312	*0.314	*0.208	*0.209
		一般	*0.065	*0.205	*0.138	*0.154	*0.318	*0.332	*0.228	*0.224
		副詞	*0.207	0.234	*0.157	*0.156	*0.315	*0.318	*0.228	*0.231
		助数詞	*0.245	*0.244	*0.153	*0.166	*0.312	*0.309	*0.219	*0.214
		形状詞	*0.243	*0.220	*0.145	*0.150	*0.318	*0.321	*0.182	*0.168
形容詞	一般		*0.241	*0.242	*0.148	*0.166	*0.308	*0.318	0.186	*0.176
形状詞	タリ		0.234	*0.242	*0.152	*0.153	0.302	*0.314	*0.177	*0.177
	一般		*0.230	*0.231	*0.139	*0.149	*0.313	*0.302	*0.214	*0.223
	助動詞語幹		0.237	*0.245	*0.141	*0.151	*0.314	*0.320	*0.203	*0.202
感動詞	フィラー		*0.243	0.235	*0.159	*0.150	*0.310	*0.310	*0.173	*0.180
	一般		*0.241	*0.241	*0.156	0.160	*0.301	*0.310	*0.169	*0.166
基準			0.235		0.160		0.303		0.186	

（* p ＜ 0.05 の場合）

析される非自立可能を今回は全て機能語として扱うことにしている。また、名詞の中の固有名詞は結合対象から除いている。

4 結果及び考察

まず、精度の基準となる値を得るために MeCab による形態素解析の結果をそのまま用いたテキストデータを fastText で学習させ得られた結果を表 2 に示す。値はサンプル数を 345 としてその中央値を求めた。

そして、対象とする品詞ごとにそれぞれ内容語に機能語を結合した場合の JWSD での分散表現の評価を行った。それぞれ 15 回ずつ評価を行い、品詞ごとのスコアの分布を調べ、その分布の中央値を表 3 に示す。統計的に有意な差が有り、分散表現で基準とする数値より相関が高いなったものを赤く表示している。

表 3 では内容語に機能語をつなげた結果、全体的に t 検定において有意な差ができることが分かった。動詞においては概ね予想通りの結果となり、名詞のサ変可能、サ変形状詞可能に機能語をつなげることで動詞の評価が UniDic 単位で分割されたときよりも良くなっていることが分かる。しかし、形容詞や副詞はほとんどの品詞で結合を行った場合に良い評価が得られるという結果になっている。機能語をまとめること自体に有効性があるのであれば、周辺の機能語が形容詞や副詞の学習に悪影響を与えることになっている可能性がある。

また、動詞の評価に注目すると副詞に非自立動詞がつながる場合が多くあったため、その分動詞の学習が困難になってしまった可能性がある。名詞の評価においても普通名詞-一般にルール (1) で機能語を結合すると評価が大きく下がっている。そのため、結合するべき品詞は厳密に選定する必要があるだろう。

今回評価に使用した JWSD に含まれる単語の語彙数や UniDic 単位で分割した際の平均単語数は表 4 のようになっている。

表 4: JWSD の統計

JWSD	名詞	動詞	形容詞	副詞
語彙数	554	1163	489	435
平均単語数	1.18	2.33	1.79	1.65

動詞は UniDic 単位ではサ変動詞などが「サ変名詞」＋「する」のように複数単語で構成されている場合が多く、本手法での貢献を最も期待していたが、結果を見ると最も改善が見られた品詞は副詞の分散表現であり、副詞に対してルール (1) を適用した場合は+0.0538 ほどの改善が見られた。今回の結果を見ると、副詞は副詞句で学習するほうが良い可能性がある。

副詞とつながる頻度の高いものでは以下のようなものがある。

- (副詞)+する
- (副詞)+お 願い いたし ます

このように機能語が続く場合はそれを全て前の内容語につなげるように処理をしているため、長い単語になることが多くなるが、いくつかの品詞に対してはそれによって分散表現の学習に有利に働いている。

5 おわりに

本研究では、単語分割の違いによって分散表現の学習にどのような影響があるかを調査し、名詞、動詞、形容詞、副詞それぞれの品詞で何かしらの条件で分散表現の改善が見られた。そのため、分散表現の学習は必ずしも細かい単位が学習に有利になるわけではないことがわかった。より良い分散表現を得るためには、内容語の品詞によって結合する品詞を変える必要があるだろう。厳密な品詞の選定を行うことでさらなる改善が見られる可能性がある。

今後としては、より良い機能語の選定やコーパスの規模を大きくした際の影響などを調べていきたいと考えている。また、現在はまだ類似語データセットによる内部的な評価しかできていないため、今後は実際に NLP のタスクでの精度にどのように影響を与えるのかを調べていきたいと考えている。

参考文献

[1] Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 298–307. Association for Computational Linguistics, 2015.

[2] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, Vol. 41, No. 4, pp. 665–695, 2015.

[3] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. Vol. 20, pp. 406–414, 01 2001.

[4] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, Vol. 8, No. 10, pp. 627–633, October 1965.

[5] Minh-Thang Luong, Richard Socher, and Christopher D. Manning. Better word representations with recursive neural networks for morphology. In *CoNLL*, Sofia, Bulgaria, 2013.

[6] Simon Baker, Roi Reichart, and Anna Korhonen. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 278–289. Association for Computational Linguistics, 2014.

[7] Billy Chiu, Anna Korhonen, and Sampo Pyysalo. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 1–6, 2016.

[8] Yuya Sakaizawa and Mamoru Komachi. Construction of a japanese word similarity dataset. *In 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*, pp. 948–951, May 2018.

[9] 猪原敬介, 内海彰. 日本語類似度・関連度データセットの作成. 言語処理学会第 24 回年次大会発表論文集, pp. 1011–1014, 2018.

[10] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析. 情報処理学会研究報告自然言語処理（NL）, Vol. 2004, No. 47, pp. 89–96, may 2004.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[13] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

[14] 吉井和輝, Nichols Eric, 中野幹生, 青野雅樹. 日本語単語ベクトルの構築とその評価. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2015, No. 4, pp. 1–8, may 2015.

[15] 押切孝将, 下平英寿. 単語分割を経由しない単語埋め込み. 言語処理学会第 23 回年次大会発表論文集, pp. 258–261, 2017.

[16] Kim Geewook, 福井一輝, 羽田哲也, 下平英寿. 単語らしい文字 n-gram の埋め込みによる単語の分散表現. 言語処理学会第 24 回年次大会発表論文集, pp. 1191–1194, 2018.