日本語動詞・形容詞類似度データセットの構築

堺澤 勇也 *　小町 守

首都大学東京

1 はじめに

近年，多くの自然言語処理 (NLP) の研究分野で単語の分散表現が活用されている．分散表現で表現されたベクトルは，固定長で密なベクトルで表され，その単語が出現する文脈が反映される．この分散表現を学習することは表現学習と呼ばれている．表現学習によって学習されたベクトルは，日本語の NLP タスクでも実績を挙げており，注目を集めている．

一般的に，表現学習により学習された分散表現は単語類似度 (word similarity) タスクや単語類推 (word analogy) タスクで評価される．これらのタスクのデータセットは，英語では用意されており，いくつかの評価セットが現存している [1, 2, 3, 4, 6, 7]．しかしながら，英語以外の言語，例えば日本語に関しては，このようなリソースは存在しない．また，これらのデータセットの多くは，名詞かつ高頻度語で構成されており，名詞以外の品詞や低頻度語を含んでいないという傾向がある．従って，それらの単語の評価を行うことはできないという問題点を持っている．

このような背景から，本稿では，日本語動詞・形容詞 [1] に関する類似度データセットを構築した．

本稿での，主な貢献を以下に示す．

- 低頻度語を含む日本語動詞・形容詞に関する類似度データセットを構築

2 関連研究

表現学習によって得られた分散表現は様々な NLP タスクで利用されている．一般的な単語の表現学習は，単語の形態素情報などを無視して，表層情報のみを利用して学習される．その結果，低頻度語や未知語に対して分散表現の学習がうまくいかないという問題が生じている．しかし，低頻度語や未知語は形態素毎に分解すると高頻度語と形態素で構成されている場合が多い（例：unkingly → un + king + ly）．このような事実から，形態素情報を活用して，低頻度語や未知語に対して適切な表現を与える研究も行われている [4, 8]．

一般的に，学習された分散表現は単語類似度タスクで評価される．英語では，WordSim353 [2]，MC [6]，RG [7]，SCWS [3] が公開されている．しかし，これらのデータセットに含まれる単語の多くは，名詞かつ高頻度語で簡単な単語であるため，名詞以外の品詞や低頻度語，未知語，形態素を持つ複雑な単語の分散表現の評価ができないという問題点があった．Luong et al. (2013) は低頻度語，未知語，形態素を持つ複雑な単語を持たないという問題を解消するため，それらを含むデータセットである Stanford Rare Word Similarity Dataset [4] を構築した．また，Baker et al. (2014) は単語類似度タスクで動詞のデータセットがないことから，WordSim353 に倣って Verb Similarity Dataset (VSD) を構築した [1].

Stanford Rare Word Similarity Dataset (RW)
Luong et al. (2013) は以下のような手順でデータセットを構築した．

単語を選出

2010 年 4 月の英語のスナップショット Wikipedia コーパスから単語を (5, 10], (10, 100], (100, 1000], (1000, 10000] の頻度で分類する．[1, 5] の頻度で出てくる単語はジャンクワードや英語でない単語が多くあったので除かれている．そして，分類した単語の中から複数の形態素を持つ単語を抽出する．これにより，前述した単語としては低頻度だが，一般的な形態素で構成されているような単語を得ることができる．抽出した単語から意味がわからないものを除くために，WordNet の synset に入っていないものは除外した．

単語ペアを構築

抽出した単語の synset の WordNet の関係（上位語 (hypernyms)，下位語 (hyponyms)，派生語 (holonyms)，部分語 (meronyms)，属性 (attributes))

*sakaizawa-yuya@ed.tmu.ac.jp

[1] 本稿では，特別に断りのないかぎり，動詞・サ変動詞を合わせて動詞，形容詞・形容動詞を合わせて形容詞と呼ぶ．

文	まさかこういった方々を対象としない、[排除する] わけではないと思いますが				
言い換えリスト	無視する	排斥する	敬遠する	除外する	排除する

図 1: 小平ら (2016) のデータセットの例

からランダムに 2 つの単語を取ってきて，1 つの単語につき 2 つの単語ペアをつくる．

類似度を人の手で付与

クラウドソーシングを利用して 10 人のネイティブなアノテータが単語ペアの類似度を 10 段階でつける．データ中にアノテータが知らない単語を含んでいる場合，それもメモとして残す．付与された類似度とメモを参考にしつつ，最終的なペアを決定する．

3 日本語動詞・形容詞類似度データセットの構築

本稿では，日本語動詞・形容詞類似度のデータセットを構築し，公開する [2]．動詞・形容詞は基本形ではなく，活用された形でも記述されているため，データセットを構築するための手順は 2.1 節で述べた RW コーパス構築の手順を参考に進めた．これは，従来の単語類似度データセットで問題点となった低頻度語がデータ中に出現しない問題を解決するためである．

今回は，小平ら (2016) [9] によって提案されている語彙平易化システムの評価データセット [3] から日本語動詞ペア・形容詞ペアを抽出する．このデータセットは，平易化の対象語を内容語（名詞，動詞，形容詞，形容動詞，副詞，サ変名詞，サ変動詞）とし，各対象語について 10 種類の文脈中での言い換えとその難易度ランキングを収録している．図 1 に小平ら (2016) のデータセットの例を示す．文中に登場する鍵括弧の中にある語が平易化の対象語として表されており，見出し語は必ずしも基本形でなく活用された形の語もある．対象語の言い換えリストが文の下の欄のように収録されている．今回はこのデータセットから以下の手順で日本語動詞・形容詞類似度データセットを構築する．

動詞・形容詞を選出

今回は，言い換え候補が動詞，サ変動詞，形容詞，形容動詞のものを抽出した．このデータでは，複数の文節にまたがるものも言い換え候補としているので，それらもすべてまとめて抽出した．構築するデータセット中に低頻度語を入れるため，抽出した動詞・形容詞を 2015 年 10 月の日本語の Wikipedia コーパスでの出現頻度（(1, 100], (100, 1000], (1000, 10000], (10000, ∞] ）毎に分類した．その結果を表 3 に示す．

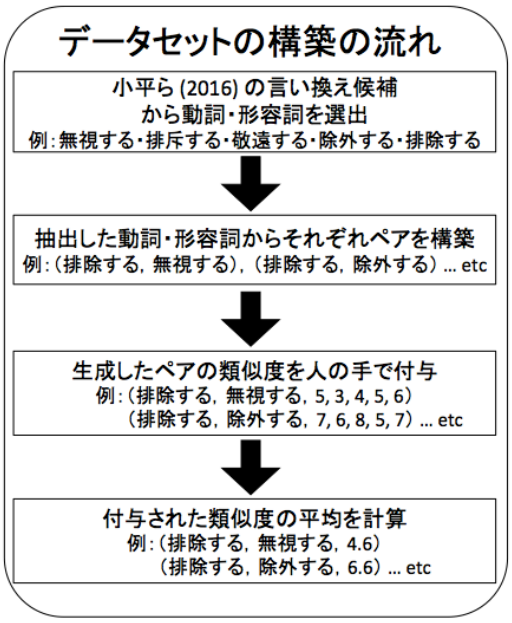

図 2: データセット構築の流れ

動詞	239 / 539 / 710 / 598
形容詞	183 / 322 / 523 / 350

図 3: 出現頻度ごとに分類した単語の数．頻度は左から，(1, 100]，(100, 1000]，(1000, 10000]，(10000, ∞]

動詞ペア・形容詞ペアを構築

抽出した動詞・形容詞にはそれに対する言い換え候補が与えられているので，それぞれを動詞ペア・形容詞ペアとして定義した．これにより，動詞 5,051 ペア，形容詞 4,033 ペアを取得した．言い換え候補の中でもペアの類似度が低いもの含まれることが確認できたので，今回は外部のリソースから類似度が低くなるようなペアを作ることはしなかった．

類似度を人の手で付与

各頻度毎に無作為にペアを抽出して，動詞 1,464 ペア，形容詞 960 ペアに対して，クラウドソーシング

[2] https://github.com/tmu-nlp/JapaneseWordSimilarityDataset

[3] https://github.com/KodairaTomonori/EvaluationDataset

単語 1	受け継ぐ	除外する	チャレンジする	しける	明白になる	迷う
単語 2	継承する	除去する	望む	あれる	反映される	止める
類似度	9.3	7.3	6	5.7	2.7	1.7
単語 1	煩わしい	望ましい	切ない	弱々しい	乏しい	甲高い
単語 2	うっとうしい	好ましい	辛い	貧しい	細かい	凄い
類似度	8.8	7.2	5.6	4	2.6	1.4

図 5: 構築されたデータセットの例

単語 1	単語 2	類似度
瞑る	つぶる	10
拭き取る	拭う	8
塞ぎ込んだ	病んだ	5
手探る	行く	2
とばせる	制御できる	0

図 4: Lancers で出した類似度の例

(Lancers[4]) を利用し，類似度を 10 段階でつけてもらい，データを収集した．アノテータは，過去の作業承認率が 95 %以上である作業者と制限した．小平ら (2016) のデータでは動詞・形容詞が入る文脈を与えてアノテーションを依頼しているが，今回の依頼では文脈は与えず動詞ペア・形容詞ペアのみを与えて類似度の付与を行った．また，VSD や RW などの単語類似度データセットはアノテーションを依頼する際に，ペアに対する類似度の例を提示していないのに対して，本稿では，アノテータに対して，類似度の例を提示した（図 4）．今回はそれぞれの動詞・形容詞ペアに対して 5 人のアノテータに類似度を付与してもらい，それらの平均をペアの類似度として定義した．

図 2 にこれまでの流れをまとめる．以上の作業で，動詞 1,464 ペア，形容詞 960 ペアの計 2,424 ペアのデータを収集した．構築された動詞のペアの例を図 5 に示す．本稿の類似度の分散の平均を計算すると，2.29 という値をとった．RW コーパスで類似度の分散の平均を計算すると 6.26 という値をとっていて，また本稿と同様の動詞の単語類似度データセットである VSD の類似度の分散の平均が 4.76 であることからも今回のデータ構築において類似度の分散は大きく改善されたことがわかる．

各ペアで類似度の分散が高かった例を以下に述べる：動詞・形容詞の関連度に起因するもの（例：速く，早くのような意味的には類似しているが，片方は時間にもう片方は速度に対して意味をなすもの：アノテータ A 10，アノテータ B 1），行為は同じだが，ムードが入るもの（例：切願する，頼む：アノテータ A 9，アノテータ B 2），ひらがなと漢字のペアを比較するもの[5]（例：ひいた，書いた：アノテータ A 7, アノテータ B 0）．今回のタスクでは，動詞・形容詞の意味的類似度を人の手でつけてもらったが，各アノテータで動詞・形容詞に対する感情やひらがなに対する漢字の想起などが異なったことからこれらのパターンで分散が大きくなったと考える．

これらは，多義語による揺れであると考えることができるので，ひとつの解決方法として小平ら (2016) の文脈を取り除くことはせず，アノテーション時に文脈を与えることで語義を一意に定めることで改善できる．また，動詞・形容詞の関連度に対する揺れは，類似度と同様に関連度に対する例をアノテーション時に与えることで抑えることができると考えられる．

日本語動詞・形容詞分散表現の評価　この節では，構築したデータセットを用いて，日本語動詞・形容詞ベクトルの評価を行う．データセットでは，動詞・形容詞は基本形ではなく，活用された形でも記述されているため，それぞれの単語ベクトルを学習した後にそれらを合成する必要がある．以下で，単語ベクトルの学習方法とその合成方法を説明する．

単語ベクトルの学習には，word2vec を利用する．単語ベクトルの学習データとして 2015 年 10 月の Wikipedia のデータ[6] 56,986,272 文を利用した．また，そのデータに MeCab 0.996 の IPADIC 2.7.0 により分かち書きの処理を施し，window 5 の階層的ソフトマックスにより単語ベクトルを学習した．ベクトルの次元は，100 次元で実験した．本稿では，学習さ

[4] http://www.lancers.jp

[5] クラウドソーシングの依頼文では,「書いた」と「かいた」のような表記のみ異なるペアの類似度は 10 になるように教唆したにも関わらず，分散が大きくなるものが見られた．

[6] http://dumps.wikimedia.org/jawiki/20151002/

データセット	結果
構築したデータセット（動詞）	18.2
構築したデータセット（形容詞）	19.0
構築したデータセット（ALL）	17.1
Luong et al. (RW) [4]	22.31
Baker et al. (VSD) [1]	64.2

図 6: 実験結果：スピアマンの順位相関係数 × 100

れた N 個の単語ベクトル $w_1, w_2, \ldots, w_N$ の平均を動詞・形容詞ベクトル v として定義する．

今回の実験では，動詞ペア・形容詞ペアの類似度は cos 類似度で計算した．生成された動詞・形容詞ベクトルで分散表現の評価実験を行う．動詞・形容詞，それらを合わせたものを対象に実験をする．実験結果と先行研究で示されている結果を合わせて図 6 に示す．

4 考察

今回は，日本語でリソースが存在しない日本語動詞・形容詞類似度データセットを構築した．単語類似度タスクでは，ランキングではなく，ペアの類似度を付与する性質から類似度の分散は大きくなる傾向にある．

WordSim353 に倣って動詞の動詞の単語類似度データセットである VSD の分散の平均は 4.76 であり，低頻度語，未知語，形態素を持つ複雑な単語を含む RW コーパスの分散の平均は 6.26 と大きな値になっている．VSD は，多義語による揺れにより，分散が大きくなっている．RW は，低頻度語や未知語のような比較的難解語を使っていることも起因していると考える．一方で，本稿でのデータセット構築では，VSD や RW と同様に多義語や低頻度語が含まれているにも関わらず，分散は 2.29 と両者に比べて大きく分散が低い結果になった．これは，タスク依頼時に類似度の例を載せることが，各アノテータの分散を抑える働きをしたと考える．しかし，感情など個人の性質による揺れまでは制限できていなかったので，より正確なアノテーションを図ることは，より細かな例を載せることによって可能になる．また，多義語によるアノテータの類似度の揺れは文脈を与えることにより，アノテーション時に語義を一意に定めることができれば改善されると考えられる．

今回は，文脈は取り除き，類似度の例を与えて，単語ペアに対して類似度を付与したが，関連度に対する例を載せることはしなかった．その結果，それによる類似度の揺れが生じてしまったため，動詞・形容詞の類似度を付与する際は，関連度に関する例や記述を与える必要があることがわかった．

構築したデータセットで，動詞・形容詞の分散表現を評価した結果，それぞれに対して大きな差はみられなかった．[4, 1] などに比べて，スピアマンの順位相関係数は低いものになっているが，先行研究はそれぞれ低頻度語や動詞に対してより適切な表現を与えるモデルを提案している．同様に，日本語の低頻度語や動詞・形容詞に対して適切な表現を与えるモデルを提案することで実験結果は改善されると考えられる．

5 おわりに

本稿では，日本語でリソースが存在しない日本語動詞・形容詞類似度データセットを構築した．また，そのデータセットを利用して日本語の動詞・形容詞の分散表現を評価した．

今後の課題としては，動詞・形容詞以外の品詞の類似度データセットの構築や単語類推タスクのデータ構築などが挙げられる．また日本語の低頻度語や動詞・形容詞に対して適切な表現を与える表現学習のモデルを構築することも考えられる．

参考文献

[1] Simon Baker, Roi Reichart, and Anna Korhonen. An Unsupervised Model for Instance Level Subcategorization Acquisition. In *EMNLP*, 2014.

[2] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing Search in Context: The Concept Revisited. In *ACM*, 2002.

[3] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving Word Representations via Global Context and Multiple Word Prototypes. In *ACL*, 2012.

[4] Minh-Thang Luong, Richard Socher, and Christopher D. Manning. Better Word Representations with Recursive Neural Networks for Morphology. In *CoNLL*, 2013.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*. 2013.

[6] George A Miller and Walter G Charles. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 1991.

[7] Herbert Rubenstein and John B. Goodenough. Contextual Correlates of Synonymy. *Commun. ACM*, 1965.

[8] Radu Soricut and Franz Och. Unsupervised Morphology Induction Using Word Embeddings. In *NAACL*, 2015.

[9] 小平知範, 梶原智之, 小町守. 均衡コーパスを用いた日本語の語彙平易化システムの評価データセット. 言語処理学会第 22 回年次大会発表論文集, 2016.