日本語類似度・関連度データセットの作成

猪原敬介（くらしき作陽大学）　内海彰（電気通信大学）

1. はじめに

意味空間モデルの性能評価には，単語類似度タスクが頻繁に用いられる。単語類似度タスクのための評価データセットは，英語では SimLex-999 [1], WordSim353 [2], MC [3], RG [4], SCWS [5], Stanford Rare Word Similarity Dataset [6], Verb Similarity Dataset (VSD) [7] など，多くが公開されている。しかし，日本語に関しては，このようなリソースは極めて少ない。著者らの知る限り，日本語動詞・形容詞類似度データセット [8]があるのみである。

近年作成されたデータセットである SimLex-999 [1]では，「類似」と「関連」の区別の必要性が指摘されている。[1]では，movie（映画）と theater（映画館）は，強く「関連」しているが（映画は映画館で鑑賞することが多いため），「類似」していない（映画館は建物という物体だが，映画は特定の形態を持つものではない）という例を挙げている。概ね，「類似」は物理的に似ていることを指し，関連は類似することを含めて，何らかの関わりがある程度を示している。

日本語における意味空間モデルの性能評価は，言語非依存の性能評価，および，日本語における応用のために重要である。本研究では，(a) 「類似」と「関連」を明確に区別した教示，(b) 性別や年代を統制した十分な数の評定者，の 2 点を特徴とする「日本語類似度・関連度データセット」の作成を試みた。

2. 方法

本研究は，(1) 調査前データセットの作成，(2) Web 調査による類似度・関連度の付与（「日本語類似度・関連度データセット」の完成），(3) 「日本語類似度・関連度データセット」の特徴について分析を行う，という手順で行われた。以下では，(1)と(2) の方法について順に述べる。

2-1. 調査前データセットの作成

できるだけ類似度と関連度が高低に広く分布したデータセットを作るために，Web 調査を行う前に，高類似・高関連／高類似・低関連／低類似・高関連／低類似・低関連という 4 条件を設定した単語ペアセットを作成した。

本研究では，類似していることを日本語 WordNet (1.1 版) [9][1]において同じ Synset（概念）に属することに，関連していることを現代日本語書き言葉均衡コーパス (BCCWJ) [10] において PPMI 類似度（共起しやすい傾向を示す指標）が高いことに，それぞれ対応すると仮定して，調査前データセットを作成した。

「類似度の高低」に基づくペア作成

まずは高類似・低類似ペアセットの作成を行った。WordNet に収録の 158,068 項目と，現代の日本語書き言葉を短単位解析するための辞書である UniDic（Ver.2.1.2）[11][2]に収録の 756,463 項目について，どちらのデータベースにも収録されている項目だけを残すと，77,280 項目（異なり語数は 27,192 語）に限定された。これは UniDic の品詞情報を付与すると同時に，複合語を除くための処理であった。

次に品詞を名詞，動詞，形容詞に限定した。名詞について，まず WordNet の品詞情報で名詞に限定すると，47,421 項目になり，さらに，UniDic の品詞情報によって普通名詞に限定すると，38,624 項目となった。このうち，異なり語数は 19,027 語であった。動詞について，WordNet の品詞情報で動詞に限定すると，19,226 語になり，さらに，

[1] http://compling.hss.ntu.edu.sg/wnja/より，「Just Japanese Words linked to Princeton WordNet Synsets」を使用

[2] https://ja.osdn.net/projects/unidic/ より，「unidic-mecab-2.1.2_src.zip」をダウンロードして，「lex.csv」を使用した。

UniDic の品詞情報によって動詞に限定すると，10,185 語となった。このうち，異なり語数は 3289 語であった。形容詞について，WordNet の品詞情報で形容詞に限定すると，8,168 語になり，UniDic の品詞情報によって形容詞に限定すると，1,904 語となった。このうち，異なり語数は 511 語であった。

次に，漢字一文字の単語は除外した。漢字一文字の単語が存在するのは名詞のみで, 全体の 9.7%であった。この操作によって, 名詞は 34872 項目，17718 異なり語となった。

さらに，BCCWJ 語彙表[3]を参考に，低頻度語を除外した。名詞について, 下位 75%未満（頻度 152 未満）を除外した。すると，12708 項目，4117 異なり語となった。動詞について, 下位 50%未満（頻度 91 未満）を除外した。すると，6894 項目，1589 異なり語となった。形容詞については単語数が少なかったため, 頻度 4 未満のものだけを除外した。すると，1904 項目，463 異なり語となった。

こうした手続きの結果，WordNet におけるそれぞれの Synset に属する単語が複数残る場合と，1 単語しか残らない場合が出てくる。同じ Synset に属する単語同士は比較的類似度が高い単語同士だと考えることができるので，高類似ペア作成用の Synset とした。一方，1 項目しか残らなかった単語は, 異なる Synset の単語とペアにすることで, 類似度が高くないペアを作ることができると考えられるので, 低類似ペア作成用の Synset とした。名詞では，前者が 8390 単語（2657 Synset），後者が 2461 単語，動詞では，前者が 5285 単語 (1654 Synset), 後者が 891 単語, 形容詞では, 前者が 1339 単語(438 Synset)，後者が 230 異なり語であった。

高類似ペア作成用の Synset に属する単語で，漢字の重複がないものについて，すべての組み合わせを作り，高類似ペア候補とした。高類似ペアについては，さらに「同じ Synset から 1 ペアのみ使用」という制限を加えた。低類似ペアについては，それぞれの品詞の低類似ペア作成用の Synset の単語をランダムにペアとして組み合わせ，漢字の重複があるものを除外した。

「関連度の高低」次元の追加

ある文脈（本研究では段落を利用した）において共起しやすい単語同士は，関連の強い単語同士だと考えることができる。そこで, BCCWJ [10] の図書館サブコーパス（短単位）から，4 単語以上を含む段落を取り出し，PPMI を実行した。ウィンドウサイズは前後 10 単語であった。上記の高類似ペアに PPMI 値を付与し, PPMI 値が上位のペアを「高類似・高関連ペア」，PPMI 値の付与できなかったペア（すなわち，共起の無かったペア）のうち，ランダムに選ばれたものを「高類似・低関連ペア」とした。上記の低類似ペアに PPMI 値を付与し，PPMI 値が上位のペアを「低類似・高関連ペア」，PPMI 値の付与できなかったペアのうち，ランダムに選ばれたものを「低類似・低関連ペア」とした。名詞ではそれぞれ 400 ペアずつ（合計 1600 ペア），動詞では 100 ペアずつ（合計 400 ペア），形容詞では 50 ペアずつ（合計 200 ペア）を選び, 調査前ペアは合計で 2200 ペアとなった。

2-2. Web 調査の手続き

予備調査

上記の 2200 ペアを 1100 ペアずつ 2 セットに分け, 各セットに対して, 2 名ずつの大学院生が「意味の分からない単語がないかどうか」を評定した。少なくとも 1 名の評定者によって「意味が分からない」とされた単語を含む 55 ペア（名詞：22 ペア，動詞：21 ペア，形容詞：12 ペア）を除外した。

最終的に残った 2145 ペアに対して，Web 調査を行った。

参加者

9253 名が Web 調査に参加した。性別・年代ごとの参加者数を表 1 に示す。

ただし，後述する手続きにおいて,「すべての評定値が同じ」「回答時間が短い参加者上位 5%」「回答時間が長い参加者上位 5%」に該当した参加者を除外することで，8132 名に減少した。

[3] http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html より,「BCCWJ_frequencylist_suw_ver1_0.zip」をダウンロードして,「BCCWJ_frequencylist_suw_ver1_0.tsv」を使用した。

表1 性別ごと，年代ごとの参加者数

	20–29歳	30–39歳	40–49歳	50–59歳	60–69歳	総計
男性	899	897	872	888	889	4445
女性	925	929	902	919	1133	4808
合計	1824	1826	1774	1807	2022	9253

評定セット

2145 ペアを，21 セットに分けた。セット間で，品詞（名詞ペア，動詞ペア，形容詞ペア）と条件（高類似・高関連ペア，高類似・低関連ペア，低類似・高関連ペア，低類似・低関連ペア）が偏らないようにした。1 セットあたりのペア数は，102 ペアもしくは 103 ペアであった。

手続き

Web 調査では，1 人の参加者が 1 セットに対して，類似度評定もしくは関連度評定を行った。そのため参加者は 42 グループに分けられ，1 グループが 1 セット・1 種類の評定を行った（例えば，あるグループはあるセットの単語ペアに対してだけ，関連度評定のみを行う）。グループ間で，参加者の性別・年代が偏らないようにした。上述の除外基準を適用後においても，1 グループには最低でも 174 名が配置されており，どの性別・年代のセル（例えば，「男性で 30-39 歳」というように，性別と年代で構成されるセル）にも最低でも 10 人が配置されていた。

調査での教示は，類似度評定条件と関連度評定条件で異なっていた。類似度評定条件では教示に加えて確認テストが行われた。類似度評定の教示では，「類似」と「関連」の区別を明確につけることが重要であることを説明し，類似に加えて関連についての説明を，例を挙げながら説明した。類似と関連の違いを示す例として，「「タイヤ」は車のパーツであるという意味で「関連」していますが，車そのものとは異なりますので，それほど「類似」はしていません。」などの教示を行った。確認テストは，この区別が付いているかを確かめる 3 択問題で，正解か不正解かのフィードバックと，解説を行った。関連度条件では，類似については触れず，こちらも例（「学者 - 書籍」の関連度は 6 (かなり関連している)，「ひき肉 - マグネット」の関連度は 2 (ほとんど関連していない) とした）を挙げながら説明した。

なお，類似度評定では 30 秒，関連度評定では 10 秒が経過しないと，説明のページから次に進むことができなかった。

その上で，「単語ペアについて，どれくらい類似（関連）しているかを，「まったく類似（関連）していない(1)」～「とてもよく類似（関連）している(7)」の 7 段階で評定し，対応する数字を選んでください。」と教示した。

その後，参加者は，1 ページあたり 10 ペアが並べられた画面で，それぞれの条件の評定を行った。

3. 結果および考察

調査前データセット作成手続きの有効性の確認

表 2 に，「日本語類似度・関連度データセット」の単語ペア例を，表 3 に，調査前に設定した条件別の評定値平均を，それぞれ示した。表 3 について，高類似の条件の類似度評定値は，低類似の条件の類似度評定値より高く，類似についての作成手続きは有効であったことが示唆される。一方，関連については，高類似・高関連が高類似・低関連よりも関連度評定値が高く，低類似・高関連も低類似・低関連よりも関連度評定値が高いので，この面では有効であったことが分かる。一方で，高類似・低関連の関連度評定が低類似・高関連の関連度評定値よりも高くなっている。これは，類似度が高いものは関連度が高くなるという性質から導かれたものだと思われる。

表2 構築した「日本語類似度・関連度データセット」の単語ペア例

品詞	名詞	動詞	形容詞	名詞	名詞	名詞
単語1	書店	着る	眩しい	加害	原型	買い物
単語2	本屋	羽織る	輝かしい	過失	石膏	同性
類似度	6.46	5.52	4.48	3.57	2.5	1.5
関連度	6.29	5.81	5.01	4.95	4.1	2.62
差	0.17	−0.29	−0.53	−1.38	−1.6	−1.12

表3 調査前に設定したそれぞれの条件における類似度・関連度評定の平均値

	高類似・高関連	高類似・低関連	低類似・高関連	低類似・低関連
類似度	4.13	3.73	2.28	1.80
関連度	5.05	4.68	3.78	2.78

これ以降では，調査前に設定した条件は考慮せず，結果として得られた類似度評定と関連度評定の関係について分析する。

類似度と関連度の関係について

全ての単語ペア（2145 ペア）について，類似度評定値と関連度評定値のピアソンの積率相関係数を算出したところ，$r = .91$ であった。類似度評定をする参加者には，「類似」の概念と「関連」の概念の違いについて教示した上で評定をしてもらったが，2145 ペアの全体的傾向としては，類似度の高低と関連度の高低は明確に連動することが分かった。

次に，類似度と関連度の評定値を 0.7 ずつの区間に分けて，クロス集計表を算出した（表 4）。表 3 の周辺度数から，類似度は評定値の低い区間に多く分布しており（1.4-2.1 の区間に最多の 753 ペアが含まれていた），関連度は評定値の高い区間に多く分布していること（4.9-5.6 の区間に最多の 497 ペアが含まれていた）が分かる。また，クロス集計表の各セルの値を見ると，類似度の全体的低さと関連度の全体的高さを反映して，対角線よりも上に多く分布していることが分かる。

数は少ないが，類似度と関連度が乖離しているペアも存在している。類似度よりも関連度の評定値が 2.0 以上高いペアは，142 ペアであった。これらは，類似度と関連度がある程度乖離している単語ペアだと解釈できるだろう。類似度よりも関連度の評定値が 3.62 高く，最大であったのは，「供給 - 需要」であった。一方，関連度よりも類似度が高い単語ペアは 10 ペアしか存在せず，その差も「死ぬ - 逝く」の 0.18 が最大と，大きいものではなかった。

表4 関連度評定値と類似度評定値のクロス集計表

		関連度評定値										
		0.0–0.7	0.7–1.4	1.4–2.1	2.1–2.8	2.8–3.5	3.5–4.2	4.2–4.9	4.9–5.6	5.6–6.3	6.3–7.0	合計
類似度評定値	0.0–0.7	0	0	0	0	0	0	0	0	0	0	0
	0.7–1.4	0	0	8	2	0	0	0	0	0	0	10
	1.4–2.1	0	0	48	353	276	55	20	1	0	0	753
	2.1–2.8	0	0	0	3	88	164	84	23	1	0	363
	2.8–3.5	0	0	0	0	0	74	150	45	8	0	277
	3.5–4.2	0	0	0	0	0	1	124	133	7	0	265
	4.2–4.9	0	0	0	0	0	0	10	226	32	0	268
	4.9–5.6	0	0	0	0	0	0	0	68	101	0	169
	5.6–6.3	0	0	0	0	0	0	0	1	38	0	39
	6.3–7.0	0	0	0	0	0	0	0	0	1	0	1
	合計	0	0	56	358	364	294	388	497	188	0	2145

注 網掛け部分が，類似度と関連度が近い単語ペアのセル。

まとめ

本研究では，「類似」と「関連」の概念の違いについて明確にした教示を行い，評定者の性別と年代について統制し，非複合語，漢字重複がないペアに限定するなどの統制を行った上で，日本語の名詞・動詞・形容詞ペアについて類似度と関連度を付与した「日本語類似度・関連度データセット」を作成した。今後，日本語の意味空間モデルの性能評価などに利用されることが期待される。

引用文献

[1] F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *Computational Linguistics*, 2016, vol. 41, pp. 665-695.

[2] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, *et al.*, "Placing search in context: The concept revisited," in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 406-414.

[3] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, 1991, vol. 6, pp. 1-28.

[4] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, 1965, vol. 8, pp. 627-633.

[5] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: 2012, Vol. 1, pp..873-882.

[6] T. Luong, R. Socher, and C. Manning, "Better Word Representations with Recursive Neural Networks for Morphology," *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, ed: Association for Computational Linguistics, 2013, pp. 104-113.

[7] S. Baker, R. Reichart, and A. Korhonen, "An Unsupervised Model for Instance Level Subcategorization Acquisition," in *EMNLP*, 2014, pp. 278-289.

[8] 堺澤勇也・小町守, "日本語動詞・形容詞類似度データセットの構築," *言語処理学会第 22 回年次大会発表論文集*, 2016, pp.262-265.

[9] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki, "Development of the Japanese WordNet," LREC, 2008, pp.2420-2423.

[10] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, *et al.*, "Balanced Corpus of Contemporary Written Japanese," *Language Resources and Evaluation*, 2014, vol. 48, pp. 345-371.

[11] 伝康晴,・小木曽智信・ 小椋秀樹・山田篤・峯松信明・内元清貴 他., "コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用," *日本語科学*, *2007*, vol. 22, pp. 101-123.