分散表現と文脈ベクトルによるオノマトペの分類の比較

古宮 嘉那子　佐々木 稔　新納 浩幸
茨城大学　工学部
{kanako.komiya.nlp,minoru.sasaki.01,hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

1　はじめに

日本語には数千ものオノマトペがあるが，それぞれのオノマトペが微妙な意味の違いをもたらしている．

Komiya ら（[1]）は文脈情報のベクトル集合をもとにオノマトペを分類し，その関係を示すためのクラスタリング手法を提案した．一方，近年分散表現による意味の表現に関する研究が盛んに行われてきている．そのため，本稿では，文脈情報の素性ベクトル集合の代わりに，分散表現のベクトルを利用して，オノマトペのクラスタリングを行い，文脈ベクトルの結果と比較，また人手で作った正解データに基づいて評価した．

2　関連研究

オノマトペとは音を表すことで感覚や感性を表す言葉であり ([5] ，[6])，近年自然言語処理の研究対象として注目を集めてきた．

本研究に最も近い研究は Komiya（[1]）である．この論文は周辺の単語やその品詞情報などの文脈情報に基づいたオノマトペのクラスタリングを提案している．これらの素性は語義曖昧性解消に使われるものであり，このシステムは語義の観点からオノマトペを分類し，可視化することを目指していた．一方，word2vec 等を利用した分散表現についての研究が近年非常によく見受けられるようになった．分散表現は，言葉の意味をベクトルで表したものであり，文脈をもとに計算され，言葉の意味の類似度の比較や，言葉の意味の合成などに利用されている．そのため，本稿では，周辺の単語やその品詞情報などの文脈情報の代わりに，この分散表現を利用して同じオノマトペの分類を行う．対象のオノマトペやコーパスを Komiya（[1]）と揃えて実験を行い，人手で作った正解データに基づいて結果を比較した．

3　オノマトペの分類

本稿のオノマトペのクラスタリングは以下の三ステップから成る．

1. それぞれのオノマトペについてコーパスから分散表現を計算する
2. オノマトペの分散表現間の距離を測る
3. オノマトペ間の距離をもとにクラスタリグを行う

コーパスには現代日本語書き言葉均衡コーパス（BCCWJ）（[2]）を利用し，形態素解析器には Mecab [1] を利用した．[2]また，word2vec[3]を利用して分散表現を作成した．この際，ウィンドウサイズは± 5 とし，分散表現の長さは 50 とした．オノマトペ間の距離は，分散表現間のコサイン距離とした．これに対し，Komiya（[1]）は形態素，品詞，かかり受け，意味コードからなる素性ベクトル集合の JSD を利用してクラスタリングを行っている．この際，ウィンドウサイズは± 2 となっている．

Komiya（[1]）と同様に，クラスタリングとしては階層型のシングルリンククラスタリングをボトムアップ方式で行った（[3]）．ボトムアップクラスタリングはそれぞれのオノマトペを表す個々のクラスタから始まり，段々と最も良く似た二つのクラスタを一つの新しいクラスタにまとめ上げていくものである．

また，シングルリンククラスタリングとは，二つのクラスタの類似度をクラスタ中の最も似ているメンバー同士の類似度とするものである．つまり，クラスタ c_u とクラスタ c_v が $c_w = c_u \cup c_v$ にマージされる際，クラスタ c_w とクラスタ c_k の類似度は，以下のように二つの個々の類似度の最大値となる．

$$sim(c_w, c_k) = max(sim(c_u, c_k), sim(c_v, c_k)) \quad (1)$$

[1]http://mecab.googlecode.com/svn/trunk/mecab/
[2]Komiya（[1]）は同じコーパスを用いているが，形態素解析器は Chasen を利用している．
[3]http://word2vec.googlecode.com/svn/trunk/

表 1: 最終的な実験におけるオノマトペの用例数の最小，最大，平均

オノマトペの種類	最小	最大	平均
照る・晴れる	17	228	80.33
寒い・冷たい	34	640	235.50
雨・雪・氷	12	1154	225.62
怒る・不機嫌・無愛想	15	5812	384.05

4 実験

実験に利用したオノマトペは，Komiya（[1]）にならい，文献 [5] の,「照る・晴れる」,「寒い・冷たい」,「雨・雪・氷」,「怒る・不機嫌・無愛想」の 4 つ分野のオノマトペのうち，コーパス中に 10 用例以上出てきているものを利用した．

つまり,「照る・晴れる」についてのオノマトペは, (1) 照る : ぎらぎら，てかてか，じりじり，ぽかぽか，かっと，かんかん，(2) 日差し : おっとり，ぽっと，(3) 晴れる : からっと，すかっと，からり，(4) 暮れる : とっぷりの 4 タイプ 12 種類である．また,「寒い・冷たい」についてのオノマトペは，(1) 寒気 : ぞくぞく，(2) 寒さ : じわじわ，しんと，りんと (3) 冷たさ : ひんやり，ひやひやの 3 タイプ 6 種類である．そして,「雨・雪・氷」についてのオノマトペは，(1) 雨 : じめじめ，どんより，ぽつぽつ，ぱらぱら，ばらばら，しょぼしょぼ，しとしと，びしょびしょ，ばらばら，さっと，ざっと，(2) 雪 : ちらほら，はらはら，(3) 雷 : ごろごろの 3 タイプ 13 種類である．また,「怒る・不機嫌・無愛想」についてのオノマトペは，(1) 怒り : むっと，むしゃくしゃ，かっと，ぐらぐら，むかむか，ぷりぷり，かんかん，がみがみ，ぷんと，ぷんぷん，ぞっと，(2) 不満 : ぐつぐつ，かちん (3) 不機嫌 : ぶすっと　つんとつんつん (4) 無愛想 : むっつり　つっけんどん (5) 態度の硬化 : きっとの 5 タイプ 19 種類である．これらのタイプは，日本語オノマトペ辞典 ([3]) に書かれている説明をもとに人手で分類した．

また，これらのオノマトペのうち,「かっと」や「からっと」のような「〜と」という形のオノマトペは，Komiya（[1]）にならい,「と」をつけた形の分散表現を利用したが，形態素解析器を変えた関係で,「りん」は「りんと」という形の分散表現が生成できなかったため,「りん」を利用した．また，表 1 は最終的な実験におけるオノマトペの用例数の最小，最大，平均を示している．形態素解析器を変えた関係で Komiya（[1]）と少々異なっている．

表 2: 分散表現と文脈表現のクラスタのエントロピー

オノマトペ	分散表現	文脈ベクトル
照る・晴れる	0.62	**0.54**
寒い・冷たい	**0.63**	0.71
雨・雪・氷	0.56	**0.54**
怒る・不機嫌・無愛想	**0.47**	0.50

表 3: 分散表現と文脈表現のクラスタの純度

オノマトペ	分散表現	文脈ベクトル
照る・晴れる	**0.67**	**0.67**
寒い・冷たい	**0.67**	0.50
雨・雪・氷	**0.79**	**0.79**
怒る・不機嫌・無愛想	**0.74**	0.68

5 結果

付録の図 1 から図 8 はそれぞれ「照る・晴れる」,「寒い・冷たい」,「雨・雪・氷」,「怒る・不機嫌・無愛想」についてのオノマトペの分散表現と文脈ベクトルの分布によるクラスタリング結果を示す．エントロピーと純度の評価に利用したクラスタ番号も併せて表示した．また，人手で作った正解データ（4 節の「タイプ」に相当）をもとに，エントロピーと純度 (文献 [4]) を評価した．表 2 と表 3 にそれぞれの結果を示す．なお，より良い結果の数値を太字で示した．

6 考察

図 1 と図 2 から，二つの手法の結果はほとんど似ていないことが分かる．しかし，図 5 と図 6，図 3 と図 4，図 7 と図 8 は，どれも最も似ているオノマトペ (「ぞくぞく」と「ひんやり」,「じめじめ」と「どんより」,「ぞっと」と「むっと」) が等しくなった．

また，表 2 から，エントロピーは「照る・晴れる」と「雨・雪・氷」のオノマトペについては文脈ベクトルの方が，それ以外に関しては分散表現の方がよいことが分かる．また，表 3 から，純度は「照る・晴れる」と「雨・雪・氷」のオノマトペについては，どちらの手法も同じであるが，それ以外については分散表現の方が優れていることが分かる．このことから，全体では，分散表現の方が正解データに近い分類を行っていることが分かった．

また，いくつかのオノマトペには多義性がある．Komiya（[1]）に引き続き，本稿でもオノマトペの語義が手に入らず，曖昧性を解消していないため，この問題の解決は今後の課題となる．

7　おわりに

本稿では，日本語のオノマトペを，分散表現をもとにクラスタリングし，先行研究であるオノマトペの文脈の分布をもとにしたクラスタリングと比較した．クラスタリング手法には階層型のシングルリンク・クラスタリングを採用し，オノマトペ間の距離にはコサイン類似度を利用した．「照る・晴れる」,「寒い・冷たい」,「雨・雪・氷」,「怒る・不機嫌・無愛想」のオノマトペの分類を行った結果,「寒い・冷たい」,「雨・雪・氷」,「怒る・不機嫌・無愛想」に関しては分散表現の結果と文脈ベクトルの分布の結果に類似が見られた．また，エントロピーと純度を人手で作成した正解データと比較したところ，分散表現の方が人手のデータと似た分け方をしていることが分かった．

謝辞

文部科学省科学研究費補助金 [若　手　B（No : 15K16046）] の助成により行われた．ここに，謹んで御礼申し上げる．

参考文献

[1] Kanako Komiya and Yoshiyuki Kotani. Classification of japanese onomatopoeias using hierarchical clustering depending on contexts. In *Proceedings of the 2011 English International Joint Conference on Computer Science and Software Engineering*, pp. 108–1013, 2011.

[2] Kikuo Maekawa. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pp. 101–102, 2008.

[3] Christopher D Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. The MIT Press, 1991.

[4] 新納浩幸. R で学ぶクラスタ解析. オーム社, 2007.

[5] 小野正弘. 日本語オノマトペ辞典. 小学館, 2007.

[6] 天沼寧. 擬音語・擬態語辞典. 東京堂出版, 1993.

付録

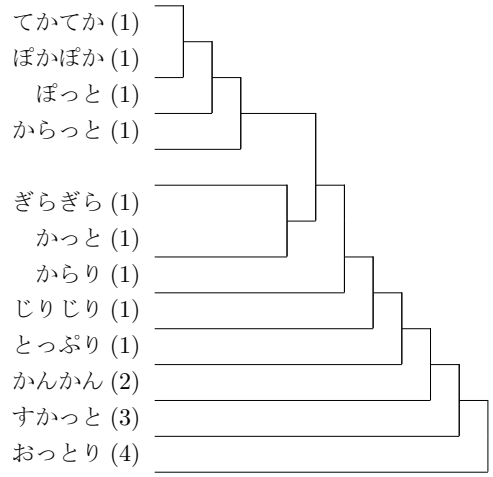

図 1: 「照る・晴れる」についてのオノマトペの分散表現による階層型クラスタリング

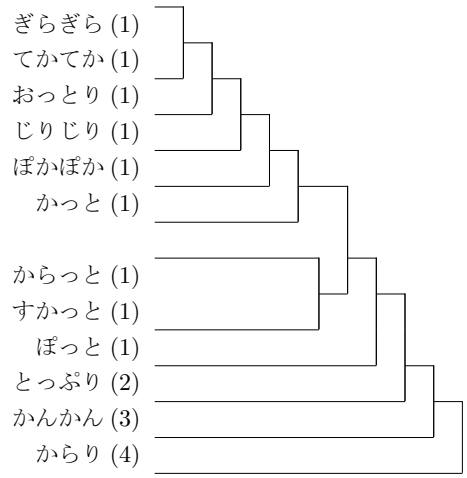

図 2: 「照る・晴れる」についてのオノマトペの文脈ベクトルの分布による階層型クラスタリング

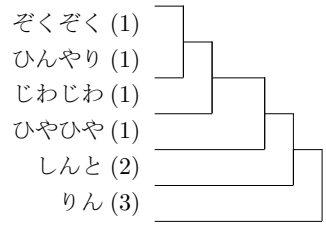

図 3: 「寒い・冷たい」についてのオノマトペの分散表現による階層型クラスタリング

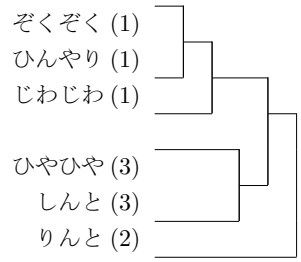

図 4: 「寒い・冷たい」についてのオノマトペの文脈ベクトルの分布による階層型クラスタリング

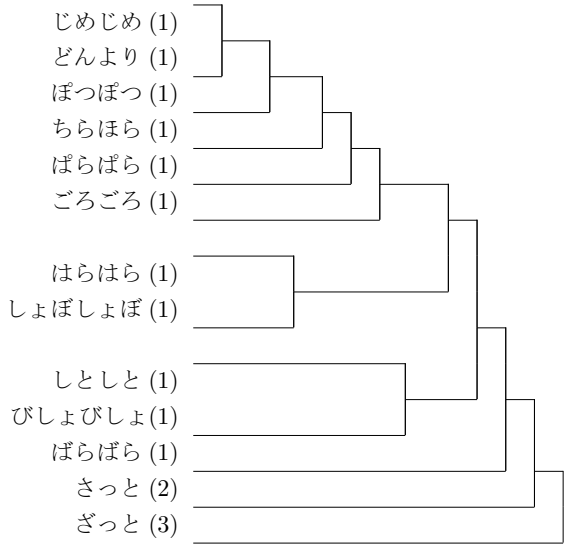

図 5: 「雨・雪・氷」についてのオノマトペの分散表現による階層型クラスタリング

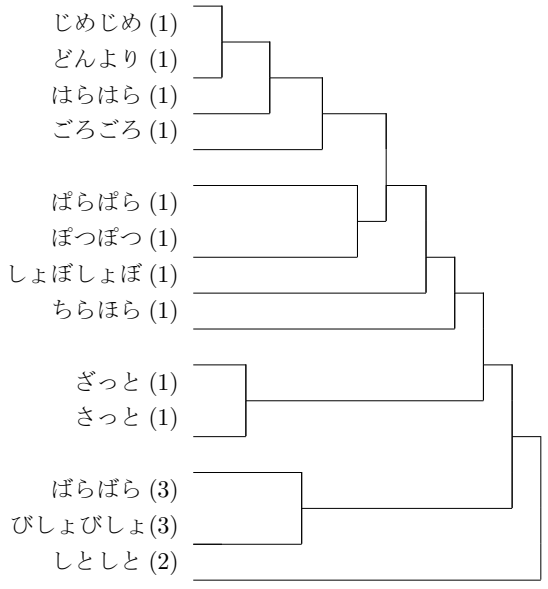

図 6: 「雨・雪・氷」についてのオノマトペの文脈ベクトルの分布による階層型クラスタリング

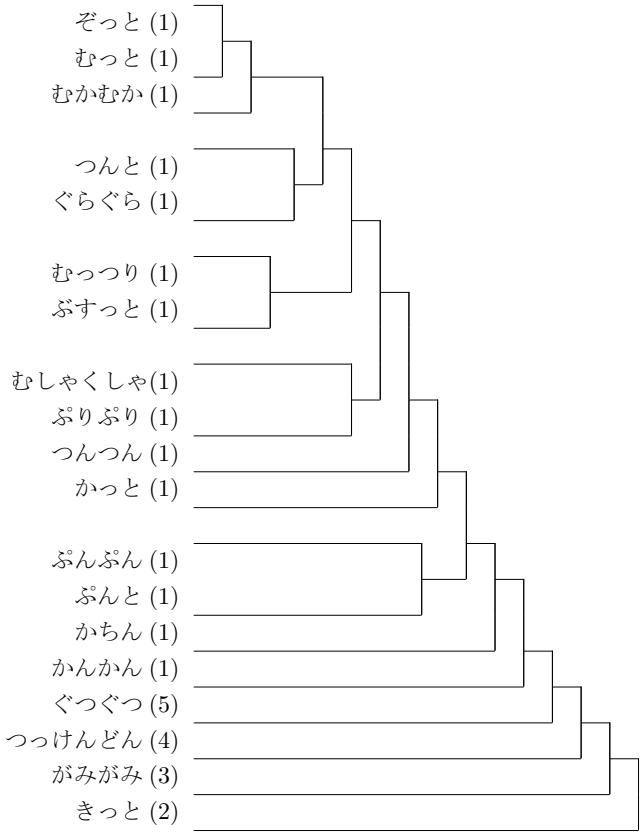

図 7: 「怒る・不機嫌・無愛想」についてのオノマトペの分散表現による階層型クラスタリング

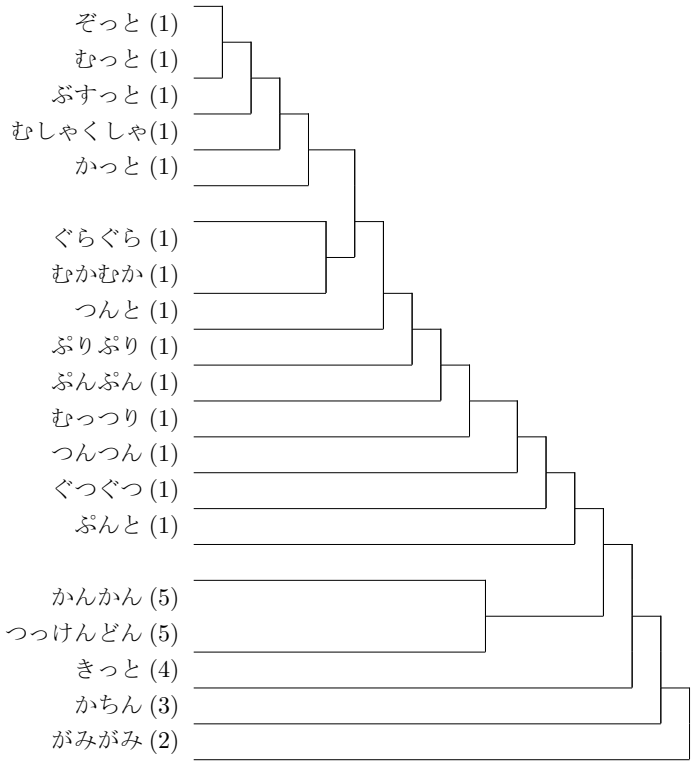

図 8: 「怒る・不機嫌・無愛想」についてのオノマトペの文脈ベクトルの分布による階層型クラスタリング