Fuzzy c-means 手法を利用したコーパスからの容器状物体の形状獲得

黒澤　義明　　竹澤　寿幸
広島市立大学大学院　情報科学研究科
{kurosawa, takezawa}@ls.info.hiroshima-cu.ac.jp

1.　はじめに

コーパス資源の蓄積は，自然言語処理研究の対象を拡大した．これに伴い，比喩等の言語現象を扱う研究も増えている．しかし，まだ不十分であると言えよう．

例えば,「鍋を食べる」という換喩の解釈としては，文字通りの解釈はありえない．鍋は食用可能な物体ではないからである．そのため，別の解釈～『容器－中身』という関係性を媒介に，“鍋（の中身）を食べる”という解釈～を要求する．

それならば，いわゆる容器として分類可能な名詞を『容器－中身』の関係性を持つ物体と考えれば，上記の解釈は容易と考えるかもしれない．しかし，実際には，山梨(1988)も指摘しているように,「押し入れをかき回す」という例文の解釈のためにはさらなる知識～容器ではない物体も『容器－中身』の関係性を持つという知識～が必要となるからである．

すなわち,いわゆる容器としての物体でなくとも，何かを入れるという機能を持つ物体は『容器－中身』の関係性を満たす可能性を有する．ここに広い意味での「容器状物体」という新しい分類が必要となる．では，こうした容器状物体の定義を如何にコンピュータに与えるのか？　単に閉空間を持つ物体と定義すればよいのか？

黒澤ら(2008)では，コーパスからの『容器－中身』の関係性記述を試みている．彼らは，“A の奥”，“A の底”等，物体 A を表す名詞とともに共起する表現に着目することを提案した．彼らの考えはシンプルであり，例えば,「鍋には深さがある一方で奥行きはさほどない．このため“鍋の底”とは言えても“鍋の奥”とは言いづらい．また，瓶は細長いため“瓶の先”と言える．しかし，鍋の場合には“鍋の先”という表現は難しい」といった人間が持つ言語直感を，コーパスから獲得し，利用することを提案したのである．

本研究も，このような共起表現（彼らが見立て詞と呼ぶ）を用いることにより，一種の言語直感のコンピュータへの構築を目指す．

ここまで述べたように，黒澤ら(2008)は上記の目的意識の上で実験を行っている．しかし，彼らの手法～Hoffman(1999)による pLSA(probabilistic Latent Semantic Analysis)を用いた次元の縮約・整理後，Kohonen(2001)の自己組織化（Self-Organizing Map：SOM）を行う手法～には問題が残る．すなわち，最終的に SOM によるクラスタリングを行っているため，複数のクラスタへの所属が困難という点である．

そこで，本研究ではソフトクラスタリングのアルゴリズム fuzzy c-means による分類を試みる．このアルゴリズムにより，名詞が持つ複数の関係性を記述できるはずである．

2.　fuzzy c-means による容器状物体の形状分類

2.1.　fuzzy c-means のアルゴリズム

本研究では，Bezdek (1981)による fuzzy c-means を用いた物体の関係性記述を試みる．以下に，アルゴリズムを記す.

2.2.　fuzzy c-means のアルゴリズム

n 個のデータを c 個のクラスタに分類するため，次の目的関数を考える．ここで帰属度 u_{ik} は $n \times c$ 行列，距離 d_{ik} は $d_{ik} = \left\| x_k - v_i \right\|^2$ を示す．x_k は k 番目のデータを，v_i は i 番目のクラスタ中心を表す.

$$J = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m d_{ik}$$

有効なクラスタの獲得（J の最小化）のため，以下の 2 式の計算（クラスタ中心及び帰属度の計算）を繰り返し行う[1]．また，パラメータ m （$m > 1$）は曖昧さを決定する．m が 1 に近づくと曖昧さは減少し，k-means 法の結果と同様になる．

[1] 導出については，宮本(1999)等，参照のこと．また，本研究は初期中心としてデータ中のランダムな 1 点を選んでいる．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

$$v_i = \sum_{k=1}^{n} (u_{ik})^m x_k \Big/ \sum_{k=1}^{n} (u_{ik})^m$$

$$u_{ik} = \left[\sum_{k=1}^{n} \left(\frac{d_{ik}}{d_{jk}} \right)^{1/m-1} \right]^{-1}$$

2.3. クラスタへの帰属確率

最終的に個々のデータの帰属度が確率で表現され，また，帰属度の和が 1 になるように正規化される．次に挙げる例（表 1）では，$c = 2$ を指定したと仮定している．

表 1　fuzzy c-means による結果例

		帰属度 (クラスタ)	
		First	*Second*
User	*A*	0.95	0.05
	B	0.65	0.35
	C	0.08	0.92
	D	0.15	0.85

このように複数のクラスタへの帰属が記述可能であるため，『容器－中身』の関連性と同時に，他の関連性についても獲得できるはずである．したがって，SOM を使用した先行研究よりも有効な結果が得られると考えられる．

3. 実験と考察

3.1. 言語データ

本研究で用いる言語データは，基本的に黒澤ら(2008，2011)で収集されたデータと同一である．その収集・加工手続きを示す．

① 容器状名詞の収集

まず，典型的な容器だけでなく，比喩的に解釈可能な名詞 A を収集する．野口(2005)の現代仮名遣い作品から，10 回以上登場する「A の中」という名詞句 265 個を収集した（ex. 頭の中）．なお，この手続きだけでは筒状の物体が少なくなるため，筒状の形状を持つ単語を追加した上で，さらに「光」・「音」・「流れ」等の抽象的，もしくは形状を有しない名詞を除き，152 語を実験材料とした[2]．

② 見立て詞毎出現率算出

池原ら(1997)の「2610 場」の下位分類から，8 語（端，角，口，奥，先，席，底，隅）を選び，容器状名詞との共起頻度を計算し，出現率にデータ変換を行った．

$$m_i\text{の出現率} = m_i \Big/ \sum_{i=1}^{n} m_i$$

3.2. 実験手続き

2 章で説明した fuzzy c-means による分類を行った．なお，パラメータ m については，1.1 から 2.0 までを変化させた．また，パラメータ c についても複数の検討を行っている．しかし，紙面の都合により，$m = 1.4$，$c = 20$ の結果についてのみ述べることとする．

3.3. 実験結果

実験により得られたクラスタの一部を示す（表 2）．ここでは，0.05 程度の値を持つ名詞のみ表示している[3]．なお，2 列目は便宜上の分類を表している[4]．

4. 考察

4.1. 全体の傾向について

黒澤ら(2008，2011)同様，基本的なクラスタリングとしては有効であると考えられる．例えば，クラスタ B においては，基本的に『文字通りの容器』が分類されている．確かに一部，容器以外の名詞(「鞄」，「ボート」) も含まれている．しかし，その形状のみから考えると，山梨(1995)に見られる＜容器のイメージスキーマ＞によって説明可能であると考えられる．

4.2. 分類の偏りについて

表 2 中の数字は，クラスタへの帰属確率を表しており，1 に近いほど，決定的な帰属性を示していると言える．したがって，1 に近い値を多数持つクラスタは，帰属の一意性が高いクラスタであると言

2 黒澤ら(2008)や黒澤ら(2011)とは異なる語が採用されている．

3 1/c 程度の意．確率の分配が等価に行われた場合，各帰属度確率は 1/c となる．したがって，何らかの意味のある値は，この 1/c を超えた場合であると考えられる．

4 容器や日用品の区別も明確ではない．あくまで，参考程度の分類であると考えて欲しい．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

える．一方，0.5 以下の値を多数持つクラスタは，他のクラスタへの帰属の可能性を残したクラスタであると言える．

表 2 では，クラスタ A 及びクラスタ B の一意性が高いと言える．半数以上のクラスタで 0.9 以上の値を有しているからである．一方，その他のクラスタでは最大値も小さく，0.5 以下の値が大多数を占める．

このように，各クラスタが持つ帰属度の値は一様ではないため，実際に使用する際（例えば，換喩の検出）には，何らかの区別が必要であることが示唆される．

4.3. ソフトクラスタ化による効果

先に述べたように，クラスタ A や B では，一意性が高いため，このようなクラスタでは，ソフトクラスタリングによる効果はあまり期待できない．その一方で，クラスタ G や F については，一定の効果が確認できる（表 3）．

クラスタ G もクラスタ F もどちらも何らかの長さを示唆させる名詞である．ただし，前者では「松林」等，より空間的に長大な長さを示している．言い換えれば，より空間的な広がりを持っていると考えられる．その一方で，後者ではホース等の長めの物体を示しており，ある種の閉鎖性を示唆していると考えられる．

この点は，表 3 に記されたコーパス中の見立て詞との共起欄（特に「奥」と「先」との共起）から確認することができる．すなわち，クラスタ G では，見立て詞「奥」との共起が大きいため，奥行きが反映された空間的な広がりを感じさせる．しかし，クラスタ F では，見立て詞「奥」との共起はほとんどなく，また，見立て詞「先」との共起がクラスタ F よりも大きいため，一点への焦点化が行われていると考えられる．したがって，より閉鎖的な印象を感じさせている．

このようなクラスタの空間的開放性-閉鎖性を利用することにより，例えば，比喩の生成～僕の未来はまるでパイプのようだ～に適用できると考えられる．比喩の理解の研究はあっても，比喩の生成の研究はなされていないため，こうした利用法は非常に有意義であると考えられる．

5. おわりに

本研究は，コーパスから容器状物質の形状抽出を試み，fuzzy c-means を使用した実験を行った．今回の実験結果は，先行研究の SOM を使用した結果に比べて，複数のクラスタへの分類～特に，人間の認知的な観点による分類～を確認しており，ソフトクラスタリングアルゴリズムの有効性を確認できたと言える．

今後の課題としては，一部のクラスタに，サブクラスタとして判定できるデータも含まれていたことから，より適切な分類尺度も考慮することが挙げられる．

謝辞

この研究の一部は，平成 22，23，24 年度広島市立大学特定研究費（一般研究）の補助を得ている．関係各位に感謝申し上げる．

参考文献

Hofmann, T. (1999). “Probabilistic Latent Semantic Indexing.” in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval”， pp.50-57.

池原悟・宮崎正弘・白井諭・横尾昭男・中岩浩巳・小倉健太郎・大山芳史・林良彦(1997) “日本語語彙大系．”岩波書店.

Kohonen，T.(2001). ”Self-Organizing Map, 3rd Edition.” 徳高平蔵，岸田悟，藤村喜久郎訳 (2005). “自己組織化マップ．” シュプリンガー・ジャパン.

工藤拓. “PLSI”, http://chasen.org/~taku/software/plsi/

黒澤義明 (to appear in 2013). “クラスタリング手法による物体形状記述と人間の印象判定～比喩・換喩の検出に向けて～．” 言語の創発と身体性，ひつじ書房.

黒澤義明，竹澤寿幸 (2011). “クラスタリング手法を利用したコーパスからの容器状物体の形状獲得．”言語処理学会年次大会.

Kurosawa, Y., Hatamoto, N., Hamada, S., and, Takezawa, T. (2012) “Comparing Clustering Algorithms for Psychomime Classification using Probabilistic Latent Semantic Analysis and Fuzzy c-Means.” In Proceedings of the 16th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems.

黒澤義明，原章，市村匠 (2008). “換喩検出を目的とした自己組織化マップ SOM による物体の形状マップ生成．”言葉と認知のメカニズム，pp.353-374，ひつじ書房.

野口英司(2005)『インターネット図書館 青空文庫』はる書房.

山梨正明(1988)『比喩と理解』東京大学出版会.

山梨正明(1995)『認知文法論』ひつじ書房.

山梨正明(2000)『認知言語学原理』くろしお出版

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

表 2　分類結果の一部

A		
檻	部屋	1.00
テント	家具	0.98
格子	家具	0.98
柵	建材	0.98
洞穴	空間	0.98
御簾	家具	0.98
押入	家具	0.98
洞	空間	0.98
村	領域	0.98
金庫	家具	0.98
城	建物	0.97
密林	自然	0.96
屋敷	建物	0.96
抽斗	家具	0.96
茂み	自然	0.96
蔵	建物	0.96
戸棚	家具	0.92
御殿	建物	0.81
戸	建材	0.76
小屋	建物	0.61
箪笥	家具	0.55
洞窟	空間	0.51
棚	家具	0.42
横穴	空間	0.27
倉庫	建物	0.23
邸	建物	0.11
荷物	日用品	0.11
ポケット	紙・布	0.09
雑木林	自然	0.08
山	自然	0.05
雑草	自然	0.05
林	自然	0.04
…	…	
sum		22.29

F		
ホース	日用品	0.90
襟	紙・布	0.89
抜け道	空間	0.87
草原	自然	0.84
パイプ	日用品	0.69
チューブ	日用品	0.44
トンネル	空間	0.35
草	自然	0.26
筒	容器	0.18
袖	紙・布	0.17
松原	自然	0.12
笛	日用品	0.12
靴	日用品	0.07
竹筒	容器	0.04
木	自然	0.04
…	…	
sum		6.73

B		
籠	家具	1.00
釜	容器	1.00
鞄	日用品	1.00
茶碗	容器	1.00
ボート	乗物	1.00
風呂	家具	0.99
樽	容器	0.99
コップ	容器	0.99
椀	容器	0.99
桶	容器	0.99
丼	容器	0.99
舟	乗物	0.99
バケツ	容器	0.99
鍋	容器	0.99
タンク	建物	0.99
湯呑	容器	0.99
湯槽	家具	0.96
炉	家具	0.95
壺	容器	0.92
グラス	容器	0.67
鉢	容器	0.48
盥	家具	0.32
缶	容器	0.17
浴槽	家具	0.12
井戸	空間	0.10
行李	家具	0.09
水筒	容器	0.09
棺	家具	0.06
皿	容器	0.05
靴	日用品	0.05
夜具	紙・布	0.04
…	…	
sum		21.59

G		
松林	自然	0.83
車	乗物	0.82
木	自然	0.71
松原	自然	0.39
草	自然	0.39
トンネル	空間	0.35
木立	自然	0.24
門	建物	0.20
袖	紙・布	0.15
筒	容器	0.14
チューブ	日用品	0.14
森	自然	0.13
笛	日用品	0.11
パイプ	日用品	0.08
靴	日用品	0.08
竹筒	容器	0.07
窓	家具	0.06
塔	建物	0.06
谷	自然	0.05
路地	建物	0.05
円筒	容器	0.05
林	自然	0.04
雑草	自然	0.04
部屋	部屋	0.04
草原	自然	0.04
…	…	
sum		6.66

C		
町	領域	0.69
店	建物	0.65
窓	家具	0.57
塔	建物	0.55
林	自然	0.46
雑草	自然	0.44
森	自然	0.38
穴	空間	0.37
門	建物	0.37
木立	自然	0.32
草むら	自然	0.20
路地	建物	0.17
山	自然	0.16
谷	自然	0.15
部屋	部屋	0.14
横穴	空間	0.10
雑木林	自然	0.10
建物	建物	0.09
牢屋	部屋	0.09
公園	領域	0.09
荷物	日用品	0.08
ポケット	紙・布	0.08
洞窟	空間	0.07
森林	自然	0.06
家	建物	0.06
壁	建材	0.06
木	自然	0.06
土蔵	建物	0.05
松原	自然	0.05
邸	建物	0.05
棚	家具	0.05
笛	日用品	0.05
庭	領域	0.05
袖	紙・布	0.05
…	…	
sum		8.68

H		
井戸	空間	0.63
浴槽	家具	0.60
棺	家具	0.44
盥	家具	0.44
缶	容器	0.43
水筒	容器	0.29
行李	家具	0.27
鉢	容器	0.25
グラス	容器	0.18
皿	容器	0.15
灰皿	容器	0.13
自動車	乗物	0.10
靴	日用品	0.09
花瓶	容器	0.08
箱	家具	0.08
夜具	紙・布	0.07
竹筒	容器	0.05
円筒	容器	0.05
筒	容器	0.04
火鉢	家具	0.04
壺	容器	0.04
監房	建物	0.04
…	…	
sum		5.60

D		
家	建物	0.72
公園	領域	0.66
壁	建材	0.66
部屋	部屋	0.52
建物	建物	0.51
庭	領域	0.22
路地	建物	0.19
牢屋	部屋	0.19
店	建物	0.14
森	自然	0.14
畑	領域	0.13
草むら	自然	0.12
ホテル	建物	0.12
窓	家具	0.11
穴	空間	0.11
天幕	家具	0.11
谷	自然	0.10
机	家具	0.10
山	自然	0.10
門	建物	0.09
木立	自然	0.09
雑草	自然	0.09
林	自然	0.08
町	領域	0.08
森林	自然	0.08
塔	建物	0.08
雑木林	自然	0.08
堂	建物	0.07
地下道	建物	0.07
書斎	部屋	0.07
ポケット	紙・布	0.07
荷物	日用品	0.07
寝床	家具	0.06
汽車	乗物	0.06
牢	部屋	0.06
松原	自然	0.05
横穴	空間	0.05
トランク	家具	0.05
笛	日用品	0.05
袖	紙・布	0.05
枠	日用品	0.05
監房	建物	0.05
邸	建物	0.05
教室	部屋	0.05
土蔵	建物	0.05
塀	建物	0.05
電車	乗物	0.05
…	…	
sum		8.40

E		
寝室	部屋	0.84
工場	建物	0.84
物置	建物	0.80
座敷	部屋	0.78
倉庫	建物	0.38
牢	部屋	0.31
土蔵	建物	0.31
堂	建物	0.30
寝床	家具	0.27
邸	建物	0.25
牢屋	部屋	0.15
書斎	部屋	0.14
小屋	建物	0.13
天幕	家具	0.12
棚	家具	0.12
荷物	日用品	0.10
ポケット	紙・布	0.10
横穴	空間	0.10
雑木林	自然	0.10
箪笥	家具	0.08
トランク	家具	0.08
山	自然	0.08
洞窟	空間	0.08
机	家具	0.07
森林	自然	0.06
病院	建物	0.06
監房	建物	0.05
草むら	自然	0.05
御殿	建物	0.05
教室	部屋	0.05
雑草	自然	0.05
戸	建材	0.05
林	自然	0.05
…	…	
sum		8.39

I		
毛布	紙・布	1.00
寝台	家具	0.99
輪	日用品	0.99
布団	紙・布	0.98
椅子	家具	0.38
ベッド	家具	0.12
夜具	紙・布	0.07
蚊帳	家具	0.06
床	建材	0.04
列車	乗物	0.04
…	…	
sum		5.33

表 3　クラスタ帰属度とコーパス中の共起

帰属クラスタ	単語	帰属度 G	帰属度 F	見立て詞(%) 席	端	角	口	奥	先	底	隅
G	松林-自然	0.83		0.00	7.55	0.00	0.00	33.96	52.83	0.00	5.66
	車-乗物	0.82		0.00	0.01	0.01	0.00	32.99	58.70	8.28	0.00
	木-自然	0.71	0.04	0.01	1.23	0.01	0.00	37.04	61.70	0.01	0.00
	松原-自然	0.39	0.12	0.00	8.33	16.67	0.00	16.67	58.33	0.00	0.00
	草-自然	0.39	0.26	0.00	13.95	0.00	0.00	16.28	67.44	2.33	0.00
	トンネル-空間	0.35	0.35	0.00	2.76	0.01	0.06	22.19	74.91	0.04	0.01
	木立-自然	0.24		0.00	0.06	0.00	0.00	50.00	49.89	0.00	0.06
F	ホース-日用品		0.9	0.00	0.25	0.00	6.53	0.06	93.16	0.00	0.00
	襟-紙・布		0.89	0.10	1.50	2.49	0.10	0.70	95.01	0.00	0.10
	抜け道-空間		0.87	0.00	0.00	12.50	0.00	0.00	87.50	0.00	0.00
	草原-自然	0.04	0.84	0.05	1.09	0.05	0.00	0.83	97.76	0.00	0.21
	パイプ-日用品	0.08	0.69	0.00	18.57	0.16	1.48	1.53	77.33	0.77	0.16
	チューブ-日用品	0.14	0.44	0.00	1.40	0.07	17.82	0.11	69.21	11.35	0.04
	トンネル-空間	0.35	0.35	0.00	2.76	0.01	0.06	22.19	74.91	0.04	0.01

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.