2007－NL－182 (6)
2007／11／19

グラフ構造に基づく同時クラスタリングを利用した動詞の属性クラスの抽出

竹内 孔一 †
岡山大学大学院自然科学研究科 †
koichi@cl.cs.okayama-u.ac.jp

Abstract
テキスト中に現れる動詞と名詞の格関係を利用して動詞のクラスタリングを行い意味的に類似の集合の構築を目指す．ここでの問題は名詞と動詞はそれぞれ多義であり，かつその語義が観測可能でないことである．本稿ではある名詞の集合と動詞の集合がなすクラスタがある1つのクラスタ (ある潜在的な意味クラスタ) に属すると考え，同時共起クラスタリングを適用する．Web の5億文のデータから獲得された格フレームデータ，ならびに毎日新聞 91 から 98 年を利用して得られた動詞集合について評価することで，本手法により大量のデータがあれば質の良い動詞集合が得られることを明らかにする．

Constrution of Semantic Verb Class Using Graph-Based Co-clustering Approach

Koichi Takeuchi
Graduate School of Natural Science and Technology, Okayama University.

Abstract
This paper presents our ongoing research for clustering Japanese verbs for constructing Japanese verb lexicon which is founded on the theory of lexical conceptual structure (LCS). The key issue of this research is how to extract a core cluster of Japanese verbs with a highly relating cluster of nouns because not only verbs but also nouns are polysemouns words. In this paper we applied an approach of co-clustering on the basis of graph sctructure into clustering task of verbs and nouns, and present experimental results on Japanese Verb-Case-Noun data from both large Web corpus and Maichini news paper corpus from 1991 to 1998.

1 背景

含意関係の計算や言い換えに役立つことを目標として，語彙概念構造の考え方を基に階層的分類に整理した動詞の意味辞書を人手で構築している [9]．動詞の意味辞書では各動詞は語義を最小単位として動詞間に共通する概念に対して動詞の集合を作成し階層的に分類している．よってこの核となる集合を発見することが意味辞書の拡張につながる．具体的の例を示すと「悪化」「急変」「暗転」は状態変化・悪化という同じ集合に分類させれる動詞の例である．

- 「事態/財政が悪化する」
- 「事態/財政が急変する」
- 「事態/財政が暗転する」

これらは意味としては全く同じではないが「悪い方向に変化する」という共通の概念を持ち，結果として似ている名詞の集合「事態」「財政」などと格関係として取る．こうした動詞の集合をテキストからクラスタリングを利用して取り出すのが本研究の目標である．

一方，テキストから語に関する性質をクラスタリングを利用して抽出する試みは，様々な手法が提案

されている．述語項構造における動詞と名詞の共起確率を主に利用したもの [2]，クラスタの重心を利用したもの [5]，クラスタリングを中心となる概念を潜在的に存在すると仮定する方法 [10][12] [7]，半教師ありでクラスタリングを行う方法 [6][8] など多種の方法によりある概念に近い語の集合の獲得を目指している．

しかしながら上記のクラスタリング手法では本稿の目的に合わない．理由は下記に示すように動詞も名詞も多義性が存在し，ある動詞集合に対してある名詞集合の組がある共通の概念でクラスタ化される必要があるからである．図 1 に格関係を伴い述語項

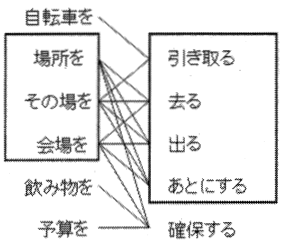

図 1: 名詞集合と動詞集合のクラスタの例

関係になる名詞集合と動詞集合の理想化した例を示している．動詞「引き取る」は「その場」など起点をあらわす名詞から去る意味があり，同様の名詞に対して「去る」や「出る」とグループをなす．一方で,「自転車/ガラクタを引き取る」のように「受け取る」という意味では別の名詞集合と密接な共起関係になる．また，動詞「確保する」はヲ格に対して「場所」や「会場」だけでなく「飲み物」など重なっているが異なる名詞集合との共起関係が密接であるため別のクラスタに分類されるべきである．こうした名詞の語義を分類語彙表など名詞の属性を分類した辞書を参考に名詞の意味を固定して動詞のみをクラスタリングすることも考えられるが，(1) 属性を入れて抽象化することでクラスタリングが不正確になること，(2) 属性も多義であり結局多義性解消が必要であることから，名詞と動詞を同時にクラスタリングが必要である．

これに対して Kurihara et al.[4] や相澤ら [11] は名詞の集合と動詞の集合，あるいは，形容詞の集合と名詞の集合というように複数の集合の組があるクラスタに属することを仮定した co-clustering という方法を提案している[1]．Kurihara et al.[4] では共起頻度に基づき relational model [3] を拡張してクラスタに対する帰属確率を求める．一方で相澤ら [11] はグラフ理論と共起頻度からなる集合対のクラスタの利得を情報量によって計算し，クラスタとして意味があるもののみ残すという方法でクラスタを獲得する．本稿では第一段階として相澤らの同時共起クラスタリング [1] を具現化してクラスタリングを行う．

以下ではまず，相澤らのクラスタリング手法について述べ，新聞記事データから格フレームをどのように獲得したかについて明らかにする．次に同時クラスタリングを行った結果について現段階での評価と展望を述べる．

2 同時共起クラスタリング

2.1 概要

詳細な手法は Aizawa[1] に譲るが，大きな枠組みとして以下のようになる．動詞 t と名詞 (格関係付き)d との共起をグラフにあらわしたとき，クラスタ S_T(動詞の集合) と S_D(名詞の集合) を認定した場合の方がクラスタとして認定しな場合よりも，全体のグラフ構造の情報量が上回った場合，クラスタとして選ぶというものである (図 2 参照)．その情報量の

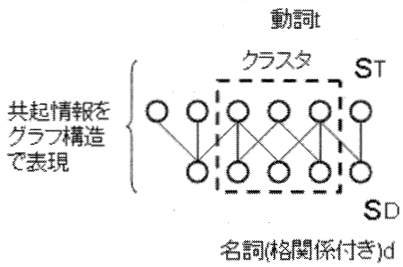

図 2: グラフ構造からクラスタを抽出する様子

評価式 δL は下記の通りで，これが正の場合のクラスタ S_T が求めたい動詞の集合になる．

$$\delta L(S_T, S_D)$$

[1] どちらも多重にクラスタに属することを許すソフトクラスタリングである．

$$
\begin{aligned}
&= L'(T,D) - L(T,D) \\
&= P(S_T,S_D) log \frac{P(S_T,S_D)}{P(S_T)P(S_D)} \\
&\quad - \sum_{t_i \in S_T} \sum_{d_j \in S_D} P(t_i,d_j) log \frac{P(t_i,d_j)}{P(t_i)P(d_j)} \quad (1)
\end{aligned}
$$

式 (1) では $L(T,D)$ が全動詞集合 T と名詞集合 D の情報量を示し，L' はクラスタ S_T と S_D を作成した場合の全体の情報量を示す．$P(S_T,S_D)$ はクラスタの出現確率，$P(t_i,d_j)$ はある動詞 t_i がある名詞 d_j と出現する確率である．この手法は同時に 2 つの要素だけでなく理論的には 3 つ 4 つと n 個の要素について同時にクラスタリングが可能である [1].

本研究では図 2 に例示したように動詞と名詞+格助詞付きのクラスタリングを行う．これは動詞のクラスタリングを行うため，格の違いにより動詞が別の分類になることを避けること，さらに，格助詞によって名詞の動詞に対する役割がかなり限定されるためより動詞の分類に役立つことがあげられる[2].

2.2 クラスタ作成の方法

前節ではあるクラスタ候補が獲得すべきものかどうかの判定方法を述べた．この節ではクラスタ候補を作成する手順を述べて，クラスタ獲得に対して自由度があることを明らかにする．以下にクラスタ作成の手順を記述する．

1. 動詞 t_i からリンクのある名詞 (格助詞付き (以下同様)) の集合を獲得する
2. **1.** で獲得できた名詞 d_j からリンクのある動詞集合を獲得する
3. リンクが多すぎるものは削除する．このとき残った動詞集合を S_T，名詞集合を S_D とする．
4. 各動詞 t_j，名詞 d_j の有効度を式 (2) (3) で評価．
5. 上記 **4.** の低いものから順に選び (1) を計算してより (1) の値が大きくなれば消す (繰り返し)．
6. 残ったクラスタ候補の (1) を計算し正ならば残し他は捨てる．
7. すべての動詞について手順 **1.** から **6.** を繰り返す．

ここで手続き **3.** ではリンクが多い，つまり動詞や名詞の特定の意味に対する使われ方に関与しない語 (例えば動詞ならば「する」や「なる」など) を排除することで信頼性の高いクラスタを得ること目指している．

$$
\begin{aligned}
&\delta L(t_i,S_D) \\
&\quad = \sum_{d_j \in S_D} P(t_i,d_j) log \frac{P(t_i,d_j)}{P(t_i)P(d_j)} \quad (2) \\
&\delta L(S_T,d_j) \\
&\quad = \sum_{t_i \in S_T} P(t_i,d_j) log \frac{P(t_i,d_j)}{P(t_i)P(d_j)} \quad (3)
\end{aligned}
$$

上記の手順でポイントとなるのは繰り返しによりクラスタ候補から式 (2)(3) を利用して不要な要素を削除する手順 **4.** と **5.** である．あるときに動詞 t_j を消去するのか名詞 d_j を消去するのかの順によって異なるクラスタを得る可能性がある．現在は名詞から消す設定を行っている．さらに，まだ試していないがある動詞のペアが同じクラスタに属することがわかっていれば，積極的にそれらの動詞を残すことでクラスタ生成をコントロールすることが可能である．この部分に関しては今後多くの実験を行うことで明らかにしたい．

3 クラスタリング実験

3.1 格フレームデータの準備

クラスタリングによる動詞分類実験を行う．必要とするデータは格関係付きの名詞と動詞の共起情報である．まず手に入るデータとして下記の 2 つを実験対象にした．

(a) Web 上の 5 億文コーパスからの格フレーム
(b) 毎日新聞 91 年から 98 年版を係り受け解析した結果得られた格フレーム

(a) は河原ら [13] によって作成されたもので構文解析器 KNP を利用して得られた格フレームである．一方 (b) は cabocha により得られた格フレームである．ただし名詞と動詞の関係において格助詞以外の関係は排除した．また能動態のみを取り出した．表 1 に格フレームデータの統計量を示す．左から格フレームの名前，名詞-助詞-動詞の種類の数，動詞の種類数，名詞の種類数を示している．毎日新聞は 91 年から一年ずつ加えて最後は 91 年から 98 年まですべてのコーパスを加えた場合のデータを作成した．表 1

[2] 例えばヲ格は状態変化を含意する動詞に対する対象であることが多い．

表 1: 格フレームデータの統計量 (種類)

	共起	動詞	名詞
Web	13651603	29213	381931
毎日 91	821954	11732	40814
91-92	1627263	13272	55677
91-93	2379097	14198	66820
91-94	3184029	15160	79488
91-95	3931152	15796	89191
91-96	4721324	16305	99569
91-97	5515753	16763	109282
91-98	5913381	16965	113831

表 2: 得られたクラスタの評価 (20 クラス (左側))

	有効	要素	purity	クラスタ数
Web	16/20	94	56.9	485
毎日 91	8/20	56	33.9	1644
91-92	7/20	50	32.0	1626
91-93	11/20	66	37.9	1416
91-94	11/20	64	34.4	1323
91-95	11/20	61	39.3	1273
91-96	12/20	51	41.2	1221
91-97	10/20	69	34.8	1173
91-98	11/20	66	36.4	1149

から Web の格フレームは新聞記事 8 年分よりも約 2 倍以上共起の種類数が多いことがわかる．ここで注目すべき特性は，動詞対名詞の種類の比率の差が大きいことである．

Web 動詞 : 名詞 = 1 : 13.1

毎日 91-98 動詞 : 名詞 = 1 : 6.7

元のコーパスの量が異なるので簡単には比較できないが，Web の方が動詞に対して出てくる名詞の種類が 2 倍多いことがわかる．つまり新聞記事の方では動詞と名詞の組み合わせがやや固定されており，動詞の語義に対しても偏った使われ方をしていることがうかがえる[3]．これは後のクラスタリングの結果にも影響を与える (次節参照)．

また係り受け解析器が KNP と Cabocha の格助詞の定義の違いにより，格助詞の数に違いが見られた．Web の方では 44 種類であるのに対して，毎日新聞では 74 種類獲得された[4]．つまり新聞記事の方が基本的に動詞と名詞の種類が少ないうえに，格助詞の種類が多いため組み合わせとして事例がより少ないことがわかる．

3.2 実験結果と考察

上記の格フレームデータに対して「動詞」対「名詞+格助詞」でクラスタリングを行った結果を表 2 に示す．表 2 の一番右側の列は獲得できたクラスタ数である．左から 2 列目から 4 列目はクラスタの評価として獲得できた動詞集合に着目して構築中の LCS 辞書と比較し，ランダムに取り出した 20 クラスタについて評価を行ったものである．左から 2 列目は各クラスタの動詞集合内で LCS 辞書の分類と比較して 1 ペア以上正しい動詞集合があれば有効なクラスタとして数え，全体してどの程度有効なクラスタが得られたかを示している．左から 3 番目の列は 20 クラスタで分類された動詞の総数である．左から 4 番目の列は 20 クラスタで測定した purity である．この数値は獲得できたクラスタの精度をあらわしている．各クラスタで複数の分類が混在しているときは最も数の多い分類を正解として数えた．

まず注目するのは獲得されたクラスタ数である．毎日新聞のデータを多くすればするほど獲得できるクラスタ数が減り，Web データにおいては最もクラスタ数が少なくなっている．しかしながら獲得できたクラスタ内の動詞数をみると表 2 の 3 列目に示すようにデータを増やすほど動詞が多くなる．この性質は本手法の特徴なのか調整によって変わるのかは現在調査中で明らかではない．しかしながら，表の 4 列目 purity に示すように大量のコーパスから得られた Web データに対する動詞のクラスタリングの結果が，新聞記事 8 年分のデータに対する結果より精度が良いことが示されている．また 2 列目の有効クラスタ数を比較しても Web データの方が優れている．

この原因が単純に格フレームの元となるコーパスの量に依存するのか，動詞と名詞の種類の豊富さによるのかはさらに他のコーパスに対して実験を行わ

[3]実際，我々の研究室では現在人手により新聞記事に対して動詞の語義を付与する作業を行っているが，新聞記事ではある特定の語義が頻出して均一的に語義が得られず例文獲得に苦労している．

[4]例えば「とかいう」「として」など複合格助詞に関する登録が多く見受けられた．このあたりの整理は次回の課題である．

なければ明らかにできないが，新聞記事のデータ量を変化させた場合，単純に精度の向上が見られるわけでない部分が動詞と名詞の組み合わせの質に関連しているように見受けられる．

次に，具体的にどのようなクラスタが得られたか例を表3に示す．

表3: 格フレームから得られたクラスタの例

	Web データ
動詞	打ってある, 振ってある, ふってある, うってある
名詞	ふりがな-が, 杭-が, 振り仮名-が
動詞	伏する, ふす, 付す
名詞	毘-に, 一笑-に, 不問-に
動詞	疑う なくす 保つ うしなう 失う 取り戻す
名詞	行き場 ヲ格 正気 ヲ格
	毎日新聞8年分
動詞	発表, 速報
名詞	棋譜-を 逮捕-と
動詞	対戦, 対局, 挑戦
名詞	段-と 名人-と
動詞	走る 向ける 移動 陥没 通行 開通 間に合う
名詞	開幕-に 道路-が

表3ではWebデータの方が毎日新聞8年分に比べてほとんど言い換えに近い動詞集合が得られていることがわかる．つまり，現段階の設定ではデータ数が多くなればなるほどとても緊密な動詞と名詞のクラスタ，つまり，言い換えが可能なものがよくとれるという結果が得られているようである．これはクラスタの評価に対して式(1)が相対的にクラスタの良さを評価することが関連している．しかし，この結果はLCS辞書の拡張という点からは良い結果ではない．ここで目標としている動詞集合は共通する概念を共有する集合であり，それは言い換えよりも少しゆるい集合である．例えば表3における「発表」と「速報」はLCSの分類[5]では「あるものを伝える」という意味を含意するとして同じ分類に属している[6]．よって表2のように不要な要素がどれだけ少ないかを中心に評価するとWebデータの方が優れているが，とれた内容についてより重要なものがどちらかといわれるといくつかの事例をみるかぎり，新聞記事のデータも有効である例が見受けられる．これは結局，全体で求めたいクラスタに対してどの程度の解を得ているかというrecallを直接的に求められないところから半分の性質しか評価されていないこと，さらに比較しているクラスタの要素が全入力に対して一部しかないため，単純なrecallで比較できないことが原因である．これに関連して多重のクラスタに同じ分類が入っている場合の評価も今は行われていない．これらの特徴を考慮して評価法を再考するのは今後の課題である．

表3の具体例で他に気づく点として基本的な動詞の取り出しの問題がある．Webの格フレームデータは既に動詞が切り出されているが「てある」といった助動詞を含むと格関係が変わってしまうため正確に動詞の性質を切り出せなくなる可能性がある．大量のテキストデータを係り受け解析するのは容易ではないが大量かつ本研究に適した格フレームデータを獲得する必要がある．

また表3にも現れているが「なくす」と「取り戻す」のように反意語が同じクラスタとして獲得できる．LCS辞書では反意語もできるかぎり記述するため必要である．反意語は共通概念があり，その方向性が違うということで反意語と認識されると仮定できるので，本来同じクラスタに属することは誤りではない．しかしながら，反意語としてクラスタが出来るのではなく似ている意味の語と同じクラスタとして獲得されることから，これらをさらに分類する手法について検討したい．

4 まとめと今後の課題

新聞記事コーパスの格フレームデータ，Webの格フレームデータを利用して概念を共有する動詞クラスタリングの結果について分析した．その結果，表層の格関係のみを利用しているだけにもかかわらず言い換え関係に近い動詞のグループが名詞のグループと同時に獲得できる見通しを得ることが出来た．さらに反意語関係も混在するが獲得データに観測されるので，今後この分類と評価について検討を加えたい．

現段階ではAizawa(2002)の共起情報に基づくク

[5]現段階で7473事例，動詞4425語に対して分類している．
[6]単語ではなく語義単位で分類されている．

ラスタリング手法のみを試しているが，今後ベクトル法による多重の特徴量を利用した場合とのクラスタリング，さらに少数の正解データを与えた場合の半教師あり学習を行い精度の比較を行いたい．コーパスとして最終的には国立国語研究所が現在作成している日本語書き言葉コーパス[7]を利用してコーパスの質と量の違いによる動詞の属性クラスタの獲得の違いについても明らかにする予定である．

謝辞

本研究は文部科学省科学研究費補助金「特定領域研究」「代表性を有する日本語書き言葉コーパスの構築:21 世紀の日本語研究の基盤整備」(代表:前川喜久雄) の援助を受けた．また，毎日新聞社様には新聞記事の利用を許諾いただいた．記して深く感謝する．

参考文献

[1] Aizawa, A.: A method of Cluster-Based Indexing of Textual Data, *Proceedings of COLING 2002*, pp. 1–7 (2002).

[2] Hindle, D.: Noun Classification from Predicate-argument Structures, *Proceedings of 28th Annual Meeting of the Assoication for Computational Linguistics*, pp. 268–275 (1990).

[3] Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. and Ueda, N.: Learning Systems of Concepts with an Infinite Relational Model, *Proceedings of the 21st National Conference on Artificial Intelligence*, pp. 381–388 (2006).

[4] Kurihara, K., Kameya, Y. and Sato, T.: Discovering Concepts from Word Co-occurrences with a Relational Model, 人工知能学会論文誌, Vol. 22, No. 2, pp. 218–226 (2007).

[5] Pereira, F. and Tishby, N.: Distributional Clustering of English Words, *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*, pp. 183–190 (1993).

[6] Wagstaff, K., Cardie, C., Rogers, S. and Schroedle, S.: Constrained K-means Clustering with Background Knowledge, *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 577–584 (2001).

[7] Wunsch, H. and Hinrichs, E. W.: Latent Semantic Clustering of German Verbs with Treebak Data, *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pp. 151–162 (2006).

[8] 杉山一成, 奥村学: 多義語の曖昧性解消への半教師有りクラスタリングの適用, 特定領域研究「日本語コーパス」平成 19 年度全体会議予稿集, pp. 115–120 (2007).

[9] 乾健太郎, 藤田篤, 竹内孔一: 含意関係計算のための事態オントロジーの開発に向けて, 信学技法, *NLC2006-89*, pp. 13–18 (2007).

[10] 萩原正人, 小川泰弘, 外山勝彦: シソーラス自動構築における PLSI の利用, 情報処理学会, 自然言語処理研究会, 2005-NL-166, pp. 71–78 (2005).

[11] 相澤彰子, 中渡瀬秀一: 係り受け関係を利用した類語・例文辞書構築法と大規模コーパスへの適用, *Proceedings of the 20th annual conference of the Japanese Society for Aritificial Intelligence* (2007). 2E1-5.

[12] 持橋大地, 松本裕治: 意味の確率的表現, 情報処理学会, 自然言語処理研究会, 2002-NL-147, pp. 77–84 (2002).

[13] 河原大輔, 黒橋禎夫: 高性能計算環境を用いた Web からの大規模格フレーム構築, 情報処理学会自然言語処理研究会, 2006-NL, pp. 67–73 (2006).

[7] http://www.tokuteicorpus.jp/