項関係にある名詞との共起を考慮した動詞のクラスタリング

真野光平 †，竹内孔一 ††
† 岡山大学工学部情報工学科，†† 岡山大学大学院自然科学研究科
†kohei@cl.cs.okayama-u.ac.jp，††koichi@cl.cs.okayama-u.ac.jp

1　はじめに

テキスト中に現れる動詞と名詞の共起関係を利用して，動詞のクラスタリングを行い，意味的に類似した動詞集合の構築を目指す．その結果得た動詞集合を，語彙概念の考え方を基に階層的分類に整理した動詞の意味辞書 (以下 LCS 辞書)[6] の内容拡張に役立てることが目的である．LCS 辞書では，動詞が語義を最小単位として，動詞間に共通する概念により動詞集合を作成し，階層的に分類している．その共通の概念を持つ動詞集合をクラスタリングにより自動構築できれば，LCS 辞書の拡張をより容易に行うことが可能になる．例えば，この意味辞書において，動詞「進行する」「はかどる」は，進行状況の変化・進歩・進行という同じ分類に属する動詞集合の例である．

- 「計画/研究が進行する」
- 「計画/研究がはかどる」

この 2 つの動詞は，完全に同じ意味ではないが,「次の段階へと進んで行く」という共通の概念を持ち,「計画-が」「研究-が」といった共通の名詞と格関係を取っている．このような動詞集合を動詞と名詞，格の共起関係から構築することを目指している．

テキストから語の性質をクラスタリングにより抽出する試みとして，様々な手法が提案されている．[3, 4, 5, 7, 9] しかしながら，動詞のクラスタリングにおいては，動詞と名詞に多義性があるため，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．つまり，先程の「進行する」の例でいえば,「進行する」が「バス-が」と共起する場合は,「はかどる」と同じ意味分類ではなくなり,「移動する」といった動詞とクラスタが生成されるべきである．このような動詞と名詞の共起を考慮して，多義性の問題を解消しようとすると，名詞と動詞を同時にクラスタリングする必要がある．

このような動詞と名詞を同時にクラスタリングする手法として，Kurihara et al.[2] や相澤ら [8] は，名詞の集合と動詞の集合のような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．本論文では，まず相澤らのクラスタリングを実装し，その結果を考察していく．また，文書単語間に潜在的な共通概念を仮定し，確率モデルを用いて文書単語行列を低次元に圧縮する PLSI との精度の比較も行う．

2　クラスタリング手法

2.1　同時共起クラスタリング

概要 詳細な手法は Aizawa[1] に譲り，ここでは概略のみ説明を行う．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い場合よりも，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 1 参照)．情報量の評価式 δL は下記の通りで，

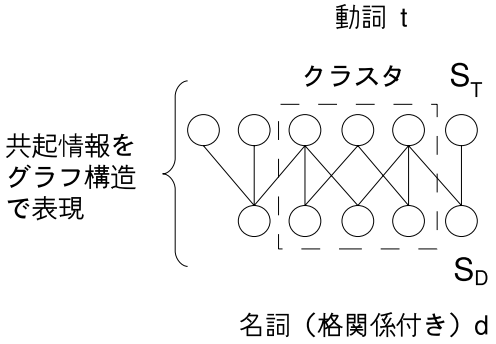

図 1: グラフ構造からクラスタを抽出する様子

これが正となるクラスタ S_T を求めたい動詞の集合とする．

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

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．

本研究では図 1 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において，格の違いにより動詞が別の分類になることを避け，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と考えられるからである．

クラスタ作成方法 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について記述する．

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_i，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．

$$\delta L(t_i, S_D) = \sum_{d_j \in S_D} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)} \quad (2)$$

$$\delta L(S_T, d_j) = \sum_{t_i \in S_T} P(t_i, d_j) \log \frac{P(t_i, d_j)}{P(t_i)P(d_j)} \quad (3)$$

2.2 PLSI

動詞の分類を獲得する方法として萩原ら [7] は，PLSI の利用を提案している．ここでは比較のために，萩原らと同様の手法を利用してある動詞に類似した動詞を得ることで提案手法との比較を行う．ここで，動詞 w が潜在意味 z を介して，名詞（格助詞付き）d と共起すると仮定する．PLSI 学習パッケージ[1]を利用し，動詞 w が z に属する確率 $P(z|w)$ を計算する．潜在クラス z は 100 個と仮定している．動詞 w_1，w_2 に対する確率ベクトル $P(z|w_1)$，$P(z|w_2)$ を用い，動詞間の類似度を，α=0.99 とした (4) 式の Skew Divergence を用いて計算した．p,q は確率分布であり，それぞれ $P(z|w_1)$，$P(z|w_2)$ である．

$$S_\alpha(p||q) = KL(p||\alpha q + (1-\alpha)p) \quad (4)$$
$$KL(p||q) = \sum_x p(x) \log(p(x)/q(x))$$

その後各動詞に対して，類似度が近い動詞でグループを作った．これは，ある動詞に対する類似度の距離が小さいグループを集めただけで，実際にクラスタリングは行っていない．しかし，PLSI の精度を評価する際の参考になると考える．

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．下記の 3 つのデータに対して実験を行った．

(a) Yahoo!知恵袋 45725 組の質問とベストアンサーを係り受け解析して得られた格フレーム
(b) 毎日新聞 91 年から 98 年版を係り受け解析して得られた格フレーム
(c) Web 上の 5 億文コーパスからの格フレーム

(a)(b) は CaboCha の係受け解析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を排除し，能動態のみ取り出した．「走り／始める」といった形態素に分割される複合動詞は前に掛かる動詞，つまり「走る」のみを取り出している．一方，(c) は河原ら [10] 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．(a)(b) と (c) の間には，構文解析器や抽出ルールの違いがあるため，それが次章以降に述べる実験結果に影響を与えている．

[1] http://chasen.org/~taku/software/plsi/

表 1 に格フレームデータの統計量を示す．統計量としては，Yahoo!知恵袋が最も少ない．毎日新聞 91-98 年に比べて，Web コーパスの方が，2 倍以上の格フレームデータであることが分かる．

表 1: 格フレームデータの統計量 (種類)

	共起	動詞	名詞
Yahoo!	227985	8034	27405
毎日 91	817920	11731	40764
91-92	1457993	13267	55619
91-93	2026784	14191	66758
91-94	2633561	15149	79417
91-95	3191462	15784	89111
91-96	3773427	16294	99487
91-97	4349547	16755	109185
91-98	4627037	16956	113731
Web	9608226	13700	343119

3.2 実験条件

本実験では上記の格フレームデータ対して，以下の 3 つの条件で実験を行う．

(a) 共起頻度を考慮し，同時クラスタリング
(b) 共起頻度を全て 1 とみなし，同時クラスタリング
(c) 共起頻度を考慮し，PLSI

共起頻度とは格フレームデータにおいて，格関係付きの名詞と動詞が実際に共起した回数のことである．いずれも，共起頻度が 1 度のみの関係は，係受け解析の誤りもあり信頼性がない．システムに与える入力としては除外した．

同時クラスタリングにおいて 2 つの条件下で実験を行う理由は，事前実験で，(a) の場合，生成クラスタ数が少ないという事象が確認できたからである．適用した同時クラスタリングで，共起頻度の値が離れた動詞が選択されなかったため，このような結果が出たと考えられる．よって，(a)，(b) 二つの条件を用いて，結果の比較を行う．

3.3 実験結果の評価方法

生成されたクラスタ全てに対して，人手で評価することは困難であるので，出力されたクラスタの先頭から 20 個を評価の対象とする．また，有効数と purity という二つの値を用いて評価を行う．

有効数とは，既存の LCS 辞書の分類と比較を行った結果，20 クラスタ中有効と考えられる要素を含むクラスタの割合を示している．

purity[11] とは，20 クラスタ中の各クラスタにおける有効な要素数の割合に全体要素の重みを考慮して計算した値であり，クラスタの精度を表わす指標となっている．

4 実験結果と考察

4.1 評価値による考察

前述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．共起頻度を考慮した結果を表 2 に，全て 1 回出現とみなした際の結果を表 3 に示す．

表の右から二番目のクラスタ数とは，クラスタリングにより生成されたクラスタの個数を表している．ま

表 2: 頻度を考慮した同時クラスタリングの評価

	有効	要素	purity	クラスタ数	平均要素
Yahoo!	7/20	43	0.349	534	2.36
毎日 91	7/20	46	0.304	1636	2.51
91-92	9/20	44	0.432	1665	2.75
91-93	10/20	55	0.400	1553	2.91
91-94	10/20	63	0.333	1439	3.01
91-95	7/20	54	0.315	1462	3.13
91-96	8/20	49	0.408	1374	3.22
91-97	7/20	68	0.309	1315	3.33
91-98	10/20	68	0.397	1286	3.29
Web	14/20	115	0.417	651	5.25

表 3: 頻度を 1 とした同時クラスタリングの評価

	有効	要素	purity	クラスタ数	平均要素
Yahoo!	10/20	79	0.342	1242	3.20
毎日 91	7/20	47	0.319	3854	2.51
91-92	5/20	42	0.238	4323	2.36
91-93	12/20	64	0.391	4315	2.34
91-94	10/20	46	0.478	4295	2.31
91-95	12/20	41	0.585	4165	2.24
91-96	11/20	40	0.550	4154	2.20
91-97	10/20	40	0.500	4152	2.19
91-98	10/20	40	0.500	4113	2.19
Web	10/20	40	0.500	1973	2.06

た一番右の平均要素とは，各クラスタが含む動詞の平均要素数である．全クラスタ中の動詞の要素数をクラスタ数で除算することにより算出した．

頻度を考慮した場合の結果を見ると，生成クラスタ数が，毎日新聞 93 年から統計量が増えるに連れ減少し続けている．1 クラス当たりの平均要素数自体は増えているが，全体として扱う要素の数は結果として減っていた．purity に関しては，統計量の変化による向上は十分確認できないが，有効要素の数が Web コーパスでは大幅に増えている．Web コーパスでは，次節で述べる格フレームデータの違いもあったため，purity が値として上昇しなかった．

一方頻度を 1 とした場合，統計量が最小の Yahoo!と最大の Web 以外は，生成クラスタの数が安定して，頻度を考慮した場合より多くなっている．また，データ量が増えるにつれ，purity も 0.50 周辺の安定した値を取っていて，クラスタとしての有効性は高い．1 クラスの平均要素自体は減っているが，頻度を考慮した場合に比べ，クラスタ数，精度ともよりよい結果となっている．しかし，1 クラスの平均要素が減ると，複数動詞間の関係を捉えにくくなる．今回は頻度を考慮した場合と，すべて 1 とみなした場合という両極端な値を選んで実験を行ったが，その中間的な値を用いた，頻度を一定度考慮しつつ，頻度値の差を減らすといった手法を考える必要がある．

また，PLSI の出力を評価した結果が表 4 である．要素数 2 とはある動詞とそれに最も類似度が近い 2 つ組のグループである．ただし，類似度が同じ動詞が 2 つある場合は，3 つ組で扱っているが，要素数 2 の項目で purity 等計算している．グループ生成の要素数を 2, 3 と小さな数にして評価を行ったのは，同時クラスタリングで頻度を全て 1 として扱った場合の要素数と条件を一致させるためである．グループの要素数を 5 とした場合も表に記した．

表 4: PLSI の評価

	要素数 2		要素数 3		要素数 5	
	有効	purity	有効	purity	有効	purity
Yahoo!	3/20	0.133	10/20	0.313	13/20	0.304
毎日 91	4/20	0.195	5/20	0.180	11/20	0.274
91-92	2/20	0.100	5/20	0.200	13/20	0.314
91-93	3/20	0.143	6/20	0.230	12/20	0.350
91-94	4/20	0.200	7/20	0.262	14/20	0.347
91-95	4/20	0.200	8/20	0.279	11/20	0.284
91-96	5/20	0.250	5/20	0.200	14/20	0.340
91-97	4/20	0.200	6/20	0.217	15/20	0.350
91-98	6/20	0.300	8/20	0.317	14/20	0.350
Web	5/20	0.244	7/20	0.283	10/20	0.294

要素数が 2，3 の場合は，統計量を増やすごとに徐々に安定傾向は見られるが，有効要素，purity 共に良い値とは言えない．要素数を 5 としてグループを生成した場合のスコアが最も良く，purity は 0.34 から 0.35 程度と比較的安定した値を取っている．これは，PLSI で求めた側近の類似度ではなく，それらを集めた固まりとして見た方が有効であることを示している．

また，PLSI は，実行の度に結果が変わるため，複数回実行してその平均を取ると精度が上がることが示されている [7] が，今回は一度の実行に対して，評価を行っている．今後，実際のクラスタリングを行うことで，精度が上がることも予想される．しかし，この値は両手法の比較に大いに参考になり，現状では同時クラスタリングの方が動詞グループの生成に有効であると言える．

4.2 生成クラスタの内容考察

実際に同時クラスタリングによりどのようなクラスタが得られたかを表 5 に示す．出現頻度を 1 とした場合を対象とする．

表 5: 同時クラスタリングにより得られたクラスタの例

	Web データ
動詞	たすける, 助ける
名詞	衆生-を, 強き-を, 消化-を, われ-を
動詞	思うようになる, 思ったりする
名詞	せい-と, 恨み-に, どう-と, どうか-と, 誇り-に
動詞	それる, 脱線
名詞	本題-から, はなし-が, テーマ-から, 趣旨-から, タイトル-から, 話題-から
	毎日新聞 91-98 年分
動詞	頒布, 販売
名詞	税込み-で 配給-という 実費-で
動詞	かぶさる, かかる
名詞	布団-が, 架線-に
動詞	合わせる, 併願
名詞	試験 と, 前期 と

Web データでは,「助ける」「たすける」といった言い換え表現がやや多く出力された．これは変換の有無といった自由な表記がされやすい Web 資源であるといった要因が考えられる．また,「思うようになる」,「思ったりする」といった一般的に辞書に掲載しない複合動詞を含むクラスタが多く生成されるのも特徴であった．

Web と毎日新聞の格フレームデータ構築の際の構文解析器や，取得規則が異なることが原因に挙げられる．このような「-になる」,「-たりする」といった複合動詞に関しては，同様の意味分類に当たると考えられる場合にも，不正解として評価した．これは，意味辞書構築支援においてそのような複合動詞に対応する予定がなく，かつ意味分類が困難なためである．しかし，これが，前節における Web データが統計量が多いに関わらず，クラスタリング結果の purity が向上していない要因とも考えられ，今後不要な格フレームデータを除外したデータでクラスタリングを行った結果の再評価が必要と言える．

Web の「それる」，毎日新聞 91-98 年の「かぶさる」を含むクラスタは LCS 辞書構築支援の目的では非常に有効である．これらは，名詞との共起によってそれぞれ「状態変化あり　進行状況の変化　状態不変化　脱線」,「状態変化あり　位置変化　位置関係の変化（物理）」という分類に当たる．同時クラスタリングでは実際にクラスタが生成される動機となった，名詞グループが動詞とともに出力できる．そのため，辞書構築支援の際，動詞がどの名詞との共起により扱われたか理解しやすく，作業者の効率の良い作業を可能とする．

本研究では，LCS 辞書構築支援を目的とすることから，単純な言い換え表現ではなく，また類義語よりもさらに広い意味概念でのクラスタの構築を目指している．この点では，Web 資源と新聞記事のどちらにおいても有効な事例を観測することが出来た．

PLSI の出力例を表 6 に示す．PLSI では，ある動詞が潜在意味 z に属す確率が計算されるため，同時クラスタリングのように，グループが選択された動機となった名詞との共起関係を求めることはできない．そのため，動詞のみを出力している．

表 6: PLSI により得られたクラスタの例（5 要素）

	毎日新聞 91-98 年分
動詞	さばく, 片付ける, 扱う, 追い掛ける, コントロール
動詞	担保, 没収, 買い戻す, 裏書, 保有
動詞	かくれんぼ, 挙兵, 昏睡, 買物, 歓送
	Web データ
動詞	怒り狂う, 激怒, 困惑, 落胆, がっかり
動詞	はいりこむ, 納めてある, なげこむ, コンビ, まぎれ込む
動詞	切腹, 出家, 自害, 挙式, ゴール

毎日新聞 91-98 年分を例にとると,「さばく」「片付ける」「扱う」といった良い要素を含むクラスタが生成される例もあれば，下段「かくれんぼ」を含むクラスタのように全く意味的な関連性が感じられないクラスタも生成された．後者の原因に，PLSI では，同時クラスタリングとは異なり，名詞動詞間の実際の共起がなくても，動詞，名詞の潜在クラスが確率ベクトルとして決定されてしまうことが挙げられる．

5　まとめ

本論文では，動詞の意味辞書構築支援のための名詞との共起を考慮したクラスタリングについて実験を行った．PLSI と同時クラスタリングにおいて比較を行い，名詞との実際の共起を考慮できる同時クラスタリングが共通する属性を持つ動詞のグループを獲得するという目的に対して，よりよい精度を示すことを明らかにした．今後 PLSI を複数回実行した結果の平均化，実際のクラスタ生成を行い，更に両手法の精度を比較していきたい．また，得られたクラスタが LCS 辞書に対し，どれだけカバーしているかを評価する方法を検討する．LCS 辞書のデータを正解セットとして与える半教師ありクラスタリングも考えられ，より有効な動詞集合を構築する手法を考察する．

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] 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).

[3] 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).

[4] 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).

[5] 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).

[6] 竹内孔一, 乾健太郎, 竹内奈央, 藤田篤: 意味の包含関係に基づく動詞項構造の細分類, 言語処理学会第 14 回年次大会, B5-2, (2008).

[7] 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).

[8] 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.

[9] 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).

[10] 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-171, pp. 67–73 (2006).

[11] Hotho, A., Staab, S. and Stumme, G. : WordNet improves text document clustering, *In Proc. of the SIGIR 2003 Semantic Web Workshop* (2003).

[12] 竹内孔一: 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, 情報処理学会, 自然言語処理研究会, 2007-NL-182, pp. 39–44 (2007).