2E1-5

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

Automatic extraction of synonyms with sample phrases using dependency analysis of text and its application to large-scale corpora

相澤 彰子*1　　中渡瀬 秀一*2
Akiko AIZAWA　Hidekazu NAKAWATASE

*1国立情報学研究所 / 総合研究大学院大学
National Institute of Informatics / Graduate School of Advanced Studies

*2総合研究大学院大学 *2006 年 3 月まで
Graduate School of Advanced Studies *Up to March 2006

This paper focuses on issues in automatic extraction of synonyms from large scale untagged corpora. In the paper, a coocurrence analysis-based method is first introduced where synonyms are extracted with their accompanying particle-verb pairs. Next, different schemes for word-space reduction are compared in a viewpoint of basic structural features of the bipartite graph representation of nouns and particle-verb pairs. The schemes are also compared using an evaluation set that are separately extracted from the target corpora. It is shown that a clustering-based scheme is advantageous in reducing noises caused by the semantic ambiguity of frequent particle-verb pairs.

1. はじめに

本稿では、タグなし自然言語文による大規模コーパスからの知識抽出法に焦点をあて、係り受け解析から得られる大量の語ペアから類語や例文を効率的にマイニングするための手法を検討する。

テキスト処理における類語の自動抽出では、共起頻度（確率）行列に統計尺度を適用して、類似する語をランキングする方法が多く用いられる。この場合の統計尺度には様々なバリエーションが存在するが、一般に抽出語数と抽出精度の間にはトレードオフの関係があり、質のよい結果を求めると得られる類語の数は少なくなり、逆に多くの語を網羅しようとすると類語の質は低下することになる。すなわち、抽出に用いるコーパス自体による限界が存在する。

この問題を解決するための現実的な手段は、コーパスの規模を拡大することである。近年では性能のよい構文解析ツールが利用可能となり、あらかじめタグが付与されていない自然言語テキストを対象とする場合でも、係り受け関係に基づき比較的ノイズの少ない共起情報が高速かつ大量に抽出できるようになった。たとえば Web をコーパス資源として活用する場合には、コーパスの規模は従来の新聞記事の場合と比較して少なくとも 1 桁～2 桁のオーダーで増大することになる。

ここで、コーパスの大規模化に伴う問題点として次をあげることができる。コーパスの規模拡大において、コーパス中に含まれる語の統計量は、べき則にしたがうと考えてよい。すなわち高頻度語ほど増分が大きく、拡大後のコーパスでも依然として支配的な影響を持つことが予想される。しかしながら高頻度語は、意味的な広がりを持つ場合も多く、特定性の高い語を分析する手がかりとしては不向きである。これより、コーパスの大規模化においては、高頻度語の影響を取り除き、その背後に隠れた情報に焦点をあてる処理が必要であると考えられる。ノイズ削減を目的としたこのような処理を、本稿では以下「共起情報の選択」と呼ぶことにする。

連絡先: 相澤彰子，国立情報学研究所，〒 101-8430 東京都千代田区一ツ橋 2-1-2, aizawa@nii.ac.jp

上記に基づき本稿では、類語抽出処理における共起情報選択の問題に焦点をあてて検討を行う。まず、**2.** で、本稿で用いる類語・用例自動抽出法の概要を述べる。次に、**3.** で、異なる共起情報選択法を適用した場合に得られる語空間の特徴を比較する。さらに、**4.** で、別途コーパスから自動抽出した類語ペアを用いて、簡単な評価を行った結果を報告する。

2. テキストからの類語・用例自動抽出法

2.1 処理の流れ

本稿における類語・用例抽出処理の流れを図 1 に示す。

(1) まず、テキストコーパスから（< 名詞 >、< 格助詞 >、< 動詞 >）の 3 項組を抽出し、あらかじめ定めた基準にしたがって一定数を選択する。
(2) 次に、< 名詞 > と < 格助詞-動詞 > をノードとする 2 部グラフを生成して、その性質を調べる。
(3) 最後に、同時クラスタリング法を適用して類語・用例のグループ化を行う。

以下、各々の処理について具体的に述べる。

2.2 係り受け関係の抽出

まず、対象コーパスを毎日新聞 98 年版 [毎日新聞 1998] として、形態素解析 [松本, 2000] および係り受け解析 [工藤, 2002] を適用した。次に、格助詞 < を >、< に >、< が >、< は >、< で > に注目して係り受け関係にある名詞と動詞を抜き出し、{ < 名詞 >, < 格助詞＋動詞 >} を類語・用例抽出のための共起ペアとした。以降、本稿では名詞の用法を説明的に示すための格助詞と動詞の組み合わせを便宜的に「格付動詞」と呼ぶ。

上記により得られる延べ数で 2,351,452、異なり数で 1,303,895 の共起ペアの中から、対応する名詞の異なり数が閾値 α_1 （= 3）に満たない格付動詞を削除し、さらに、対応する格付動詞の異なり数が閾値 α_2 （= 3) に満たない名詞を削除して、最終的に異なる 989,468 個の共起ペアを得た。

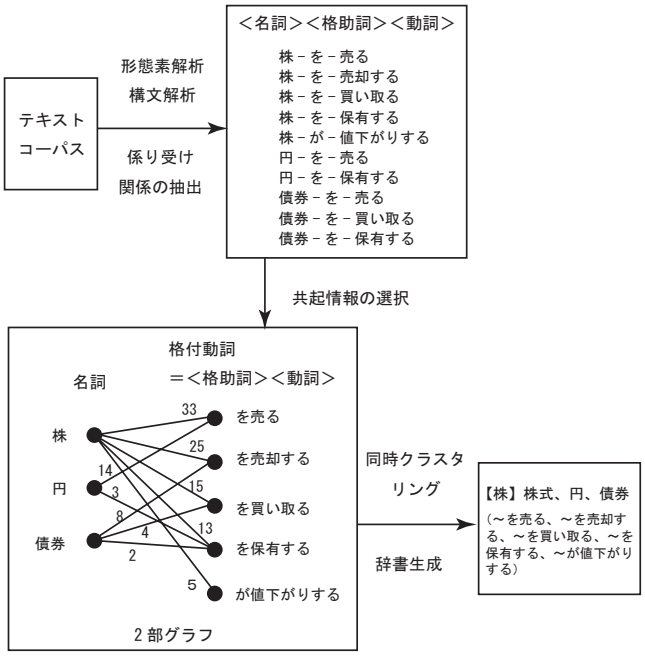


図 1: 本稿における類語・用例抽出処理の概要

2.3 共起情報の選択と２部グラフ生成

共起情報の選択法として、 [A] 頻度による動詞選択、[B] 相互情報量によるリンク選択、の２通りを適用し、両者を比較した。

[A] 頻度による動詞選択（動詞選択法）

接続する名詞の異なり数が多いもの順に、上位から指定した割合だけ格付動詞を取り除く方法。これらの格付動詞に関連した共起ペアはすべて削除されることになる。削除の対象となる格付動詞の例は、「を-する」「が-ある」「に- なる」「を-行う」「を-見る」等であった。

[B] 相互情報量による共起ペア選択（リンク選択法）

名詞と格付動詞の共起関係を相互情報量の順に並べ、その値の小さいものから順に取り除く方法。具体的にはコーパス中の名詞 $n \in N$、格付動詞 $pv \in PV$ として、各々の総出現数を $freq(n)$、$freq(pv)$、共起関係の総数を F、簡単のため $P(n) = \frac{freq(n)}{F}$、$P(pv) = \frac{freq(pv)}{F}$ として、以下で与えられる相互情報量 $M(n, pv)$ の値に基づき選択に用いた。

$$M(n, pv) = \frac{logP(n, pv)}{logP(n)logP(pv)} \tag{1}$$

相互情報量の値が小さく削除の対象となる共起ペアの例は、「今回-が-行う」「方針-に-ある」「政府-に-なる」等であった。これらは名詞・動詞それぞれは多く出現するが、組み合わせとして頻度が低かったためであると考えられる。

2.4 同時クラスタリングの適用と辞書生成

上記により選択した共起ペアに対して、クラスタリングを適用し、その過程でさらに有用な共起ペアを選択する。

[C] クラスタリングによる共起ペア選択（クラスタリング法）

情報量に基づく同時クラスタリング法 [Aizawa 2002] の適用により情報量的に結びつきの強い名詞と格付動詞のグループを同時抽出する方法。互いに近傍する領域中で関連の低い共起ペアを取り除く効果がある。

最後に、得られたクラスタを利用して、名詞グループからは類語候補を、格付動詞のグループからは対応する用例を抽出して、見出し語・類語・用例から成る辞書項目を作成した。

3. 共起情報選択の効果分析と選択法の比較

3.1 比較に用いる２部グラフ上での数量尺度

名詞の集合を N、格付動詞を PV とする。このとき抽出した共起ペアは、ノード集合 N と PV に対する２部グラフ $(N, PV, \mathcal{D})$ に対応づけて考えることができる。ただし、$\mathcal{D}$ は上記の方法により抽出された名詞と格付動詞の対応で、$\mathcal{D} \in N \times PV$ とする。

以下では、この２部グラフ上で前出の [A]、[B] ２つの選択法を比較する。具体的には、選択した共起関係の数、すなわちグラフの総リンク数を共通の変数として横軸にとり、以下の量を調べる。

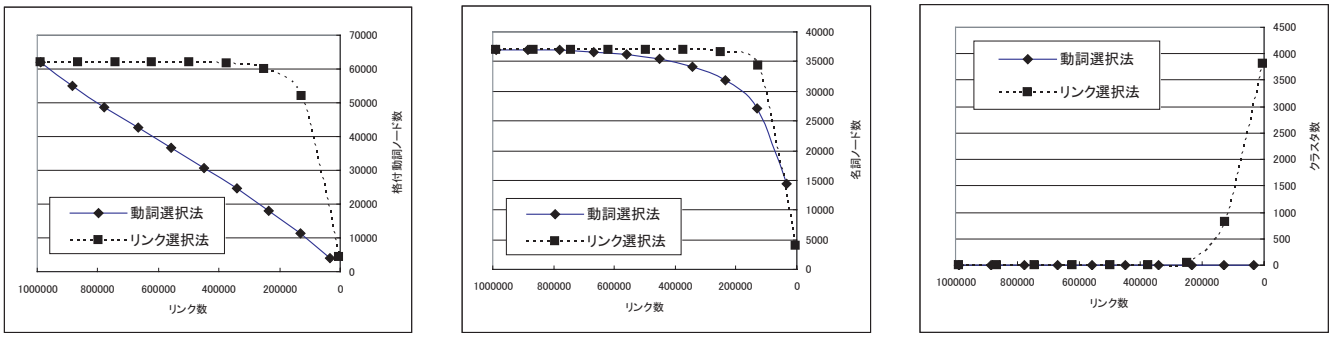
(1) ノード数
　２部グラフ上のノード数 $|N|$, $|PV|$
(2) 連結成分の数
　この値が 1 の場合、２部グラフ上のすべてのノードからすべてのノードに到る経路が存在する。
(3) 平均近傍数
　任意の $n_1 \in N$ から経路長 2 で到達可能な $n_2 \in N$ の集合を $K(n_1)$ とする。$|K(n_1)|$ を n_1 の近傍数、グラフ上のすべての n_1 に対する $|K(n_1)|$ の平均値を平均近傍数と呼ぶ。$K(n_1)$ は、名詞 n_1 と少なくとも１個の格付動詞を共有する名詞の集合であり、類語候補となる。
(4) 平均最短経路長
　２部グラフ上の２つのノード $n_1, n_2 \in N$ の間の最短経路長を $p(n_1, n_2)$ として、すべての到達可能な n_1, n_2 の組み合わせに対する $p(n_1, n_2)$ の平均値を平均最短経路長とする。
(5) 平均近傍係数
　$n_1 \in N$ の任意の２つの近傍ノード $n_2, n_3 \in K(n_1)$、n_2, n_3 と共起する格付動詞の集合をそれぞれ $PV(n_2)$、$PV(n_3)$ とする。両者の積集合と和集合の要素数の比の、n_2,n_3 のすべての組み合わせに関する平均を近傍係数 $c(n_i)$ と呼ぶ。すなわち、

$$c(n_1) = \frac{1}{|K(n_1)|(|K(n_1)| - 1)} \times \sum_{n_1, n_2 \in K(n_1), n_2 \neq n_3} \frac{|PV(n_2) \cap PV(n_3)|}{|PV(n_2) \cup PV(n_3)|} \tag{2}$$

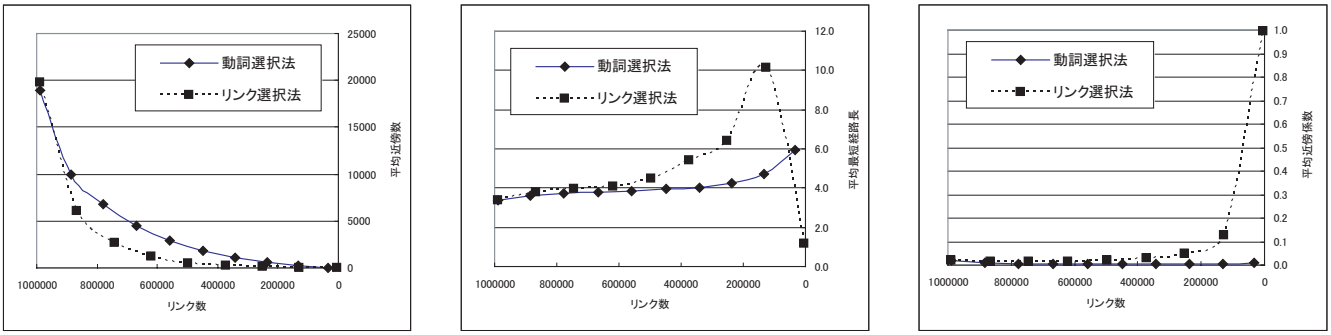
　すべての $n_1 \in N$ に対する $c(n_1)$ の平均値を平均近傍係数と呼ぶ。$c(n_i)$ はスケールフリーネットワークにおけるクラスタ係数に相当する尺度として定義するものである。近傍の名詞ノードが存在しない場合、近傍係数の値は１とする。

3.2 新聞記事１年分を用いた選択法の比較

実験では、上記の各項目について、動詞選択法とリンク選択法の影響を調べた。ただし計算時間の観点から、グラフ上のすべてのノードではなく、ランダムにサンプリングした 1000 ノードについての平均値を求めた。結果を図 2 に示す。



(a) 格付動詞ノード数　(b) 名詞ノード数　(c) 連結成分の数



(d) 平均近傍数　(e) 平均最短経路長　(f) 平均近傍係数

図 2: 2 部グラフの特徴を手がかりとした選択法の比較

まず、動詞選択法では、リンク数を減らすとノード数も減少することから、動詞選択を行っても、グラフのリンク密度は比較的高い状態のまま保たれることがわかる。このことは、平均最短経路長の変化が比較的少ないことにも表れている。また全体の 80%の格付動詞を除いても、グラフ全体はゆるく連結した状態が保たれる。次に、リンク選択法では、リンク数が減少してもノード数は大きく変化しない。一方で、グラフのリンク密度は急速に減少し、平均最短経路長は増加する。ここで一定の境界値（約 50%）を境に平均最短経路長が急速に減少するのは、それ以降はグラフの連結性が失われ、独立な小さな成分が多く生成されるためである。

このように、動詞選択法とリンク選択法は、生成する 2 部グラフの構造の上では対照的なふるまいを示し、前者では大域的特徴に、後者では局所的特徴に重点が置かれていることがわかる。なお、次節で適用するクラスタリング法は、両者の性質をあわせ持っている。

4. 類語・用例の抽出結果と分析

4.1 類語関係評価用セットの獲得

特定のコーパスにおける「類語」関係は、必ずしも人手により構築された体系的なシソーラスと対応がとれるわけではない。たとえば新聞記事コーパスの中では、「株」と「債券」は類似した文脈で出現することが多いが、分類語彙表 [分類語彙表 2004] によれば以下のように異なる分類に属する。

例 1：分類語彙表における「株」「債券」:

「株」　体 → 活動 → 経済 → 資本・金銭
「債券」　体 → 生産物 → 物品 → 貨幣・切符・証券

本稿では、「類語は必ずしも汎用的な体系によって定まるものではなく、正解はコーパスが代表する語彙空間に依存して決まる」、と考え、対象コーパスから特定の表現パターンにより抽出した名詞ペアを評価用セットとして用いることにする。

具体的には、まず、「A-や-B」という表記に注目して、{A, B} を類語関係の候補として抽出する。次に、定型的な表現を除外するため、順番を逆にした「B-や-A」の出現頻度が極端に少ないペアを削除する。さらに、コーパス中での出現頻度が A B ともに閾値 N（= 100）以上であるような 216 ペアを評価用セットとして選択する。

例 2: 評価用の類語ペアの例

(障害者, 高齢者) (テレビ, 新聞) (国, 自治体) (被害者, 遺族)
(証券会社, 銀行) (卵子, 精子)　(企業, 個人) (デモ, 集会)
(アジア, ロシア) (官僚, 政治家)　(日本, 米国) (教師, 親)

このような評価セットの構築法は、必ずしも適切な類似関係を保証するものではないが、対象コーパスの分野特性を反映する比較のための参照データとして、コーパスが大規模になった場合にも容易に適用可能であるという利点がある。

実験では次式で定める $sim(A, B)$（以下「PV 類似度」と呼ぶ）を用いて、A と B の類似の度合いを調べ、全評価ペアに対する平均を性能値とした。なお、実験に先立ち、他のいくつかの類似性尺度についても評価を行い、同様の傾向が得られることを確認している。

$$sim(A, B) = \frac{|PV(A) \cup PV(B)|}{min\{|PV(A)|, |PV(B)|\}} \quad (3)$$

4.2 評価用セットを用いた選択法の比較

図 3 は、動詞選択法とリンク選択法について、上記で定義した PV 類似度の平均値を比較したものである。本稿の評価

表 1: 「株」に関する類似ランキング上位語の比較

選択なし		動詞選択法		クラスタリング法	
株	を上場する	株	は大暴落する	株	を買わない
国債	は譲り受ける	国債	を売り抜ける	円	は買う
相場	を売り抜ける	相場	に深入りする	株式	が値下がりする
リーディングカンパニー	がわれる	同社株	を上場する	ドル	を売買する
株価	が乱高下する	不動産	を大量購入する	債券	に投資する
名義	で増やそう	株式	を強制取得する	不動産	を手放す
不動産	を引き取らせる	カメラ	を買い占める	長銀株	が売る
株式	を大量購入する	企業	を譲渡される	銀行株	が買われる
企業	で損する	銃	で増やそう	国債	を買い上げる
銃	を償還する	円	を規制される	銘柄	を所有する
損失	を売らぬ	株価	をカラ売る	株価	を譲渡する
カメラ	は割り当てられる	お金	がわれる	債権	は急騰する
円	を借り受ける	ジャンボ機	を買わされる	資産	を持たぬ
金額	を買わされる	パスタ	がだぶつく	通貨	を引き取る
絵	を平均する	偽造券	は譲り受ける	土地	は値上がりする
土地	が取引される	金魚	が乱高下する	営業権	が下げる
商品	ははね上がる	公開株	を引き取らせる	東京株式市場	を運用する
昨年	は高騰する	内野	で損する	国際優良株	を保有する

セットに関する限り、動詞選択法とリンク選択法については、選択を行わない場合がもっともよい性能値を示しており、削減による改善の効果は確認できなかった。

一方、図 3 右上に示したプロットは、クラスタリング法の適用後の 2 部グラフに関するリンク数と平均 PV 類似度の値を示している。クラスタリング法の適用によりリンク数は大幅に減少するが、PV 類似度の値は逆に向上していることがわかる。

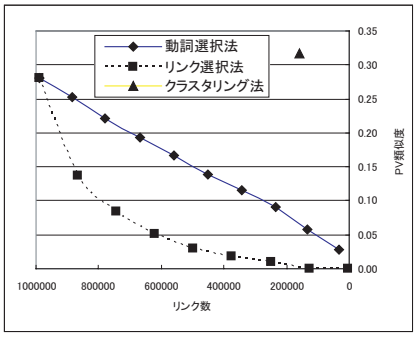


図 3: PV 類似度による比較

さらに、動詞選択法（10%の格付動詞を削除）とクラスタリング法の PV 類似度のヒストグラムによる比較を図 4 に示す。

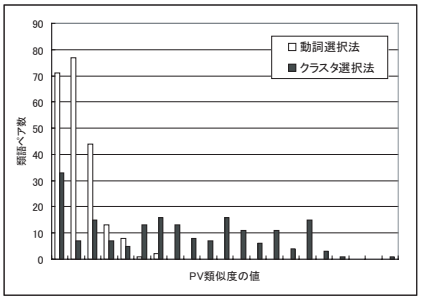


図 4: 動詞選択法とクラスタリング法のスコア分布の違い

4.3 SMART ランキングによる比較と辞書構築

表 1 は、前節の手法により選択した共起データを GETA[Geta URL] でインデックス化して、名詞・格付動詞それぞれに対して情報検索分野で一般的な重み付け尺度 SMART を用いてランキングを行った結果を示す。検索語は「株」である。たとえば「株」と「カメラ」はともに頻度が中程度の「を借り受ける」等の格付動詞と共起しているが、クラスタリング法によりこれが削除されたことで、クラスタ内の結びつきが強まっていることがわかる。最後に例 3：にクラスタリング法により作成した辞書の例を示す。

例 3: 類語・用例辞書の例

【株】
株式, 円, 株価, ドル, 国債, 銀行株, 債券, 長銀株, 不動産, 土地, 物価, 地価, 切符, 投票率, 債権, 原油価格, 平均株価, 通貨, 優先株, 価格, 銘柄, ニューヨーク株式市場, 資産, 日経平均株価, ハンセン指数, ナイフ, 証券, チケット, 国際優良株, 金融債
(～を売る。 ～を売却する。 ～を買い取る。 ～を保有する。 ～を購入する。～を手放す。 ～が売られる。 ～を買い上げる。 ～を借りる。 ～が値下がりする。 ～を引き取る。 ～を所有する。 ～を売買する。～を持たない。 ～を取得する。 ～が売り込まれる。 ～が上昇する。 ～を買い戻す。 ～が暴落する。 ～が急落する。 ～を買わない。～を引き継ぐ。 ～が買われる。 ～に投資する。 ～を発行する。 ～を預ける。 ～が売る。 ～が急上昇する。 ～は急騰する。 ～が下げる。～は下がる。 ～を持たぬ。 ～は売られる。 ～を運用する。 ～が急騰する。 ～が出回る。 ～は値上がりする。 ～は回復する。 ～を引き受ける。～を譲渡する。 ～が下落する。 ～で運用する。 ～は下落する。 ～を購入できる。 ～を証券化する。 ～を処分する。 ～を渡す。～を譲り受ける。 ～が更新する。)

5. むすび

本稿では、係り受け関係に基づき類語・用例辞書を作成するためのテキストマイニング処理について検討し、特に高頻度語に由来するノイズを低減するための共起情報の選択法について分析を行った。提案手法は、より大規模なコーパスへの適用を前提としており、今後は、さまざまなコーパスへの適用結果を比較しながら、コーパス固有用法にも取り組んで行きたい。

参考文献

[松本, 2000] 松本、北内, 山下, 平野, 松田, 高岡, 浅原: 日本語形態素解析システム『茶筌』version 2.2.1 使用説明書 (2000)

[工藤, 2002] 工藤, 松本: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, vol.43, no.6, 63-69 (2002)

[Aizawa 2002] Aizawa, A.: A Method of Cluster-Based Indexing of Textual Data, Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002) , 1-7 (2002).

[毎日新聞 1998] 毎日新聞社: 毎日新聞記事データ集 1998 年版 (1999).

[分類語彙表 2004] 国立国語研究所編: 『分類語彙表 増補改訂版』、大日本図書 (2004).

[Geta URL] http://geta.ex.nii.ac.jp/