大規模コーパスからの語義のマイニングに関する考察

田渕　史郎† 　鍛治　伸裕†† 　吉永　直樹†† 　喜連川　優††

† 東京大学大学院情報理工学系研究科　〒 113-0033 東京都文京区本郷 7-3-1
†† 東京大学生産技術研究所　〒 153-8505 東京都目黒区駒場 4-6-1
E-mail: †{tabuchi,kaji,ynaga,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　情報検索において、検索語が多義語である場合に、ユーザは意味が混在した検索結果を受け取ることになる。自然言語処理においては一般的にはこの問題は曖昧性解消の技術により解かれてきた。自然言語処理における曖昧性解消のタスクは、多義語及び、その語義が列挙された問題設定で解かれていることが多い。しかし、この問題設定であると、未知語が多義性を持っていた場合や、既知の多義語に対して辞書には登録されていない語義が生まれてきた場合に、対応できない。筆者らこの状況に対応できるように任意の語に対して語義を自動的に数え上げれるような枠組みを提案する。具体的には、任意の語に対して共起語のグラフを作り、グラフ中でクラスタリングを行うことによって、複数の共起語の集合のクラスタを求める。このクラスタは語義に対応する傾向があり、語義を自動的に数え上げることができる。

キーワード　自然言語処理, 曖昧性解消, 情報検索, グラフクラスタリング

A study on the mining of word senses from huge corpus

Shiro TABUCHI†, Nobuhiro KAJI††, Naoki YOSHINAGA††, and Masaru KITSUREGAWA††

† Graduate School of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
†† Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan
E-mail: †{tabuchi,kaji,ynaga,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　In information retrieval, a polysemous search query provides the users with a mixture of documents that are related to individual senses of the query. This problem can be solved as a kind of word sense disambiguation (WSD) in the field of natural language processing. However, since existing WSD methods assume that both words and their senses are known, we cannot directly utilize those methods to disambiguate senses of search queries that can include unknown polysemous words or words with novel senses. To tackle this problem, we present a method that determines the number of senses for an arbitrary word. We first construct a graph whose nodes are words that co-occur with the target word. We then perform a clustering of those words in the graphs to find a cluster of words whose members co-occur with each other. A preliminary analysis of the resulting clusters revealed that each cluster is likely to correspond to a single sense of the word, and the number of senses for the word can be threfore estimated from the number of clusters.

Key words　natural language processing, word sense disambigutaion, information retrieval, graph clustering

1. はじめに

近年ブログやウェブなどのテキストデータが増加しており、情報検索における多義語の問題が深刻になっている。この問題をジャガーという多義語を例に挙げて説明する。ジャガーには動物のジャガー、車のジャガー、漫画の「ピューっと吹くジャガー」のジャガーなど異なる意味がある。このジャガーをクエリとして漫画の意味で情報検索を行った場合に、他の意味のジャガーの検索結果はノイズとなってしまう。検索者に曖昧性を解消して検索結果を返すことができれば有用である。これは自然言語処理上の多義性解消の手法を用いて解くことができる。しかし、既存の多義性解消の問題は多義語およびその多義語に対する語義のセットが予め与えられていることが多く、これでは未知の語が発生した場合や、語義の数が変遷していく場合には対応できない。もし、任意の語に語義の数を数え上げれる枠組みがあれば、新語や語義の変遷に対応できる。このよう

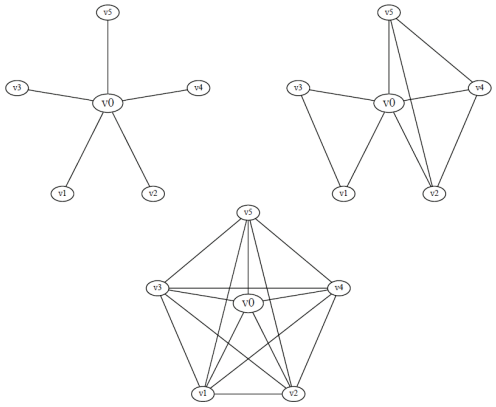

図 1　Curvature によるノード v_0 の評価

な手法について検討する。

2. 関 連 研 究

語義の自動抽出に関する研究について次の 2 つに分類して紹介する。

2.1　語の分布を利用した方法

Lin らは以下の方法で多義語の語義を抽出している [3]。まず、ターゲットとなる名詞に対して、文書 w と語 x の間の自己相互情報量を計算して、その後、文書 w と語 x との自己相互情報量を計算する。次に、語 x と語 y の類似度を計算する。こうして、例えば apple = [orange:0.7, banana = 0.5, mac:0.3, dell:0.5] のような特徴ベクトルを作りこのベクトルをクラスタリングする。そして辞書中の語を Clustering by Commity という方法でクラスタリングする。Clustering by Commity は k-means 法と似た方法であるが、優先順位付きかつ重心が一定であるという点で k-means 法と異なっている。

2.2　グラフに基づく方法

Dorow らは共起語をグラフとして表現した後に、クラスタリングを行っている [1]。無関係な複数のクラスタに属していそうなノードを曖昧な単語として以下の curvature という指標を用いて議論している。

$$curv(v) = \frac{(\#triangles\ v\ could\ participate\ in)}{(\#\ of\ triangles\ v\ articipates\ in)} \qquad (1)$$

ノード v の curvature は v を端点とするエッジだけを含む三角形の最大個数に対して実際に v が含まれている三角形の割合として計算される。図 1 の全てのグラフにおいては v_0 の次元数が 5 であり、作りうる三角形の最大個数は ${}_5C_2 = 10$ である、この個数に対して、v_0 が参加している三角形の個数は左上に関しては 0 であり、$curv(v_0) = 0$。右上については 4 個あるので $curv(v_0) = 0.4$ であり、下のグラフについては全てのノードが他のノードと連結したいわゆる極大クリークとなっているがこの場合だと $curv(v_0) = 1$ である。Curvature が低い語はグラフにおいて複数のクラスターのハブのような位置にあることが予想される。この低い語を切り離すことによってグラフをクラスタリングする。

表 1　共起語の抽出に利用した語彙統語パターン

X や Y	X も Y も	X と Y と	X,Y など

3. 提 案 手 法

本章では提案する語義の発見手法について説明する。提案手法は共起語に着目したものとなっている。多義語は、語義に応じて全く違った単語と文書中で共起する。例えば、多義語「ジャガー」には車の意味であれば、「ポルシェ」や「ベンツ」と共起するであろう。また、動物の意味であれば、「ライオン」や「シマウマ」などと共起すると思われる。したがって「ジャガー」と共起する語をうまく分類することができれば、それは語義の発見につながると考えられる。ターゲット語の共起語はウェブコーパスから語彙統語パターンを用いて抽出する。得られた共起語をノードとし、共起関係をエッジとするグラフを作る。これを**共起語グラフ**と呼ぶ。このグラフをノードをクラスタリングすることによって、語義に対応する共起語クラスタを得る。

3.1　共起語グラフの作成

共起語グラフの抽出は以下の手順で行う。

（1） 表 3.1 の語彙統語パターンを用いてコーパスからターゲット語の共起語を収集する。

（2） 共起語だけのグラフを作る。ノードは共起語を表し、エッジは共起関係を表す。この際、共起度の低いノードは語義発見の際にノイズとなるので削除する。ノード x,y 間の共起度は以下で定義される PMI を用いて求めてその値が σ より小さいノードは取り除く。

$$\mathrm{PMI}(x,y) = \log \frac{f(x,y) \times f(*,*)}{f(x,*) \times f(y,*)} \qquad (2)$$

従来手法ではターゲット語自身も含まれた共起語グラフを用いているが、本研究ではターゲット語自身は含まれない共起語グラフを用いることに注意されたい。(図 2)

3.2　共起語グラフのクラスタリング方法

3.2.1　Newman 法

もう一つのクラスタリング方法として用いたのは [2] などで用いられている Newman 法である。Newman 法は次式で定義される Q 値 (式 3) を用いたボトムアップクラスタリングの方法である。

$$Q = \sum_i (e_{ii} - a_i^2) = Tr(\mathbf{e}) - ||\mathbf{e}||^{\mathbf{2}} \qquad (3)$$

まず、全てのノードを一つのクラスタと見た初期状態を考える。そこから 2 つのクラスタを選ぶ。選ばれるクラスタはその 2 つをマージした状態の Q 値が最大になるものである。式 3 中の e_{ij} はクラスタ i とクラスタ j を結ぶエッジ数の総エッジ数に対する割合である。Q 値はこの e_{ij} のうち同じクラスタ内でエッジが張られている分だけ大きくなるが、全エッジを一クラスタにマージしないように、a_i^2 が減じられている。a_i^2 はエッジが複数のクラスタに均等に散らばっている程クコアが高くなる。例えば図 3 であると状態 N と状態 N + 1 においてマージされたことによって a_i^2 は 0.3 から 0.34 というように増えてしまう。

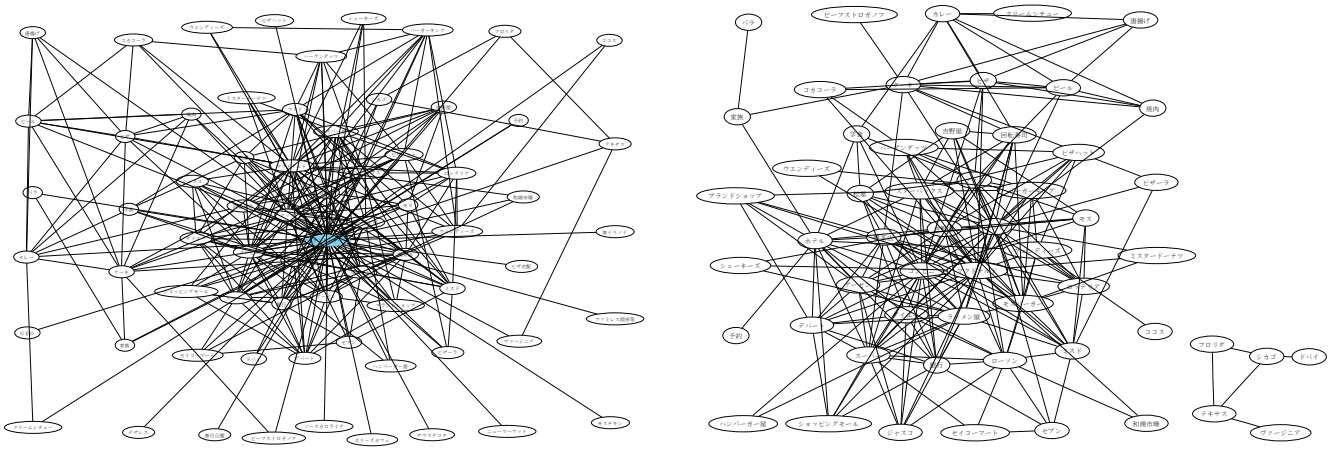

図 2　従来手法によるケンタッキーの共起語グラフ (左) と提案手法によるケンタッキーの共起語グラフ (右)

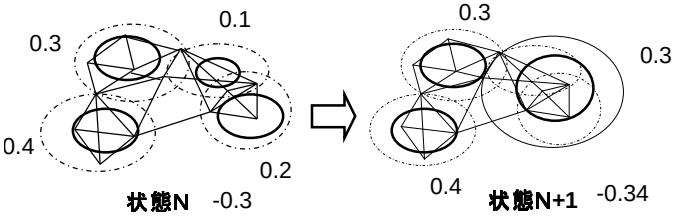

図 3　Newman 法の遷移

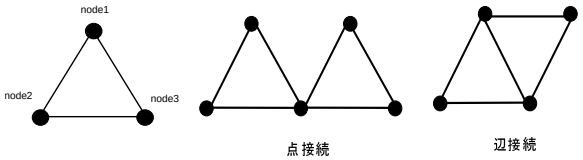

図 4　3-クリークによる最小単位 (左), 点接続と辺接続 (右)

このようにクラスタリングしすぎないように制約が掛かっている。マージングが行われるのは Q 値が増大している間、すなわち次式で計算される ΔQ が 0 より大の間である。

$$\Delta Q = 2e_{ij} - a_i^2 \tag{4}$$

この方法だと最初にクラスタ数を与えなくてよい。

3.2.2　3-クリークによるクラスタリング

一つは 3-クリークを利用したボトムアップクラスタリングである。まず、図 4 左のような 3 つのノードからなるクリーク (3-クリーク) を考える。この 3 つのノードは意味の近接性が高いと考えられる。このような 3-クリークを最小単位とし、図 4 右のような点接続でボトムアップクラスタリングを行う。具体的な手続きは Algorithm1 に示す。ボトムアップクラスタリングの結果出てきたクラスタのうち、ノード数が 3 以下のクラスタ (3-クリークそのもの) はノイズの可能性が高いので無視する。

4. 実　　験

4.1 データセット

実験は当研究室で集めた 2005 年度のウェブアーカイブ 1 億

Algorithm 1 3-クリークによるクラスタリングのアルゴリズム (点接続)

```
G = (V,E)
//3-クリークの集合の列挙　// adj(v) は v の隣接ノード集合　//
edge(u,v)⊂ E
for v ∈ V do
  for u1 ∈ adj(v) do
    for u2 ∈ adj(v) do
      if edge(u,v) exists then
        3-clique = [u1,u2,v]
        //3-クリークのインデックスを作る
        index[v] += 3-clique
        index[u1] += 3-clique
        index[u2] += 3-clique
      end if
    end for
  end for
end for
//点接続によるボトムアップクラスタリング
check[v] = null
cluster = null
//v から始めてくっつくクラスタの処理 (点接続)
for start ∈ V do
  buff := index[start]
  next if check[start] == 1 //すでに処理済み
  check[start] = 1
  do
    next if check[start] == 1 //すでに処理済み
    cluster += index[start]  // 3-クリークをクラスタに追加
    buff += index[start]
    buff = buff.uniq
    start = buff.pop
    check[start] = 1
  while buff exists
  if cluster size >= 4 then
    return cluster
  end if
end for
```

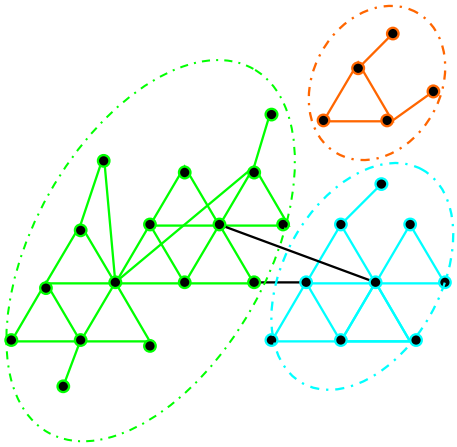

図 5　3-クリークを利用したボトムアップクラスタリング

表 2 「ロス」のクラスタリング結果

ID	クラスタに含まれる共起語	語義の判定
1	香港 ハワイ NY 日本	○（ロサンゼルス）
2	ニューヨーク シカゴ ロンドン サンフランシスコ	○（ロサンゼルス）
3	ミス コスト 手間 トラブル リスク	○（損失）
4	方向性 アフロ ニューオリズ	×
5	近藤 ボール シオン ベントン	×

7000 万文から抜き出したコーパスから多義語 74 個に対して行った。平均語義数は 2.3 個であり最大語義数は 4 個である。

4.2 評価方法

以下の手順で、求められた各クラスタに人手で正解、不正解のラベルを付与した。

（1） 共起語集合から語義を作業者が列挙する

（2） システムが求めた各クラスタが 1 で作業者が列挙した語義に対応する時、その正解のラベルをクラスタに付与する。ただし、複数のクラスタが同一の語義に対応している時一つのクラスタのみ正解とし、残りは不正解とする。

このようにして付与された正解ラベルを基にして、以下の尺度を用いて得られた結果の評価を行った。

$$\text{適合率} = \frac{\text{正解ラベルを付与されたクラスタの数}}{\text{作業者が列挙した語義の数}} \tag{5}$$

$$\text{再現率} = \frac{\text{正解ラベルを付与されたクラスタの数}}{\text{全クラスタの数}} \tag{6}$$

$$F\text{ 値} = \frac{2}{\frac{1}{\text{適合率}} + \frac{1}{\text{再現率}}} \tag{7}$$

具体例として「ロス」という多義語に対する提案手法の適合率、再現率を求めてみよう。「ロス」という単語について共起語集合から列挙された語義は地名のロサンゼルス、「損失」を意味するロスの 2 つであった。一方本手法を用いて「ロス」の共起語集合をクラスタリングした結果は表 2 のようになった。1 のクラスタは人手で地名の意味に対応しているとして正解と判定できる。2 のクラスタも地名の意味に対応しているが、地名の意味に対応するクラスタが複数存在してしまうので、不正解と判定する。3 は損失の意味のロスに正しく対応している。4,5 は正しい語義に対応しないので両方不正解とする。よって

表 3　実験結果

	適合率	再現率	F 値	平均発見語義数
Curvature	0.63	0.66	0.65	1.6
Newman 法	0.48	**0.95**	0.64	2.1
3-クリーク（$\sigma = -\infty$）	0.41	0.40	0.40	1.8
3-クリーク（$\sigma = 0$）	0.70	0.70	0.70	0.98
3-クリーク（$\sigma = 5$）	**0.74**	0.79	**0.77**	1.8

「ロス」の再現率は $\frac{2}{2} = 1.0$ と計算され、適合率は $\frac{2}{5} = 0.4$ と計算される。F 値は 0.57 になる。

4.3 実験結果

3.2.2 節、および 3.2.1 節で説明したクラスタリング方法による精度を実験により計算した。そして、ベースラインとしてそれらと 2.2 節で取り上げた curvature を用いたクラスタリング方法を実験し、比較した。ベースラインとして用いた方法では curvature の値が 0.35 以下のノードを削除するようにした。その上で残ったノードの繋がった部分を一塊とみなしてクラスタリングしている。クラスタとして出力されたもののうちでノード数が 4 以上のクラスタだけを抜き出した。

Curvature を用いたクラスタリング、Newman 法によるクラスタリング、3-クリークによるクラスタリングそれぞれの語義に関する適合率、再現率、F 値、平均獲得語義数について比較すると表 3 のようになった。3-クリーク法は PMI の複数の閾値 ($\sigma = -\infty, 0, 5$) によってエッジカットを行った上で適用した。それらの複数の結果について載せている。結果として適合率、平均語義数で 3-クリークによる方法が F 値で上回った。再現率については Newman 法が他よりもよいスコアを示した。

a）「マーチ」の例

ボトムアップクラスタリングによって

1 リングマーチ トライアンフ ＢＳＡ ミニ

2 カローラ スカイライン イスト インテグラ シビック アルテッツァ ビッツ デミオ ヴィッツ フィット シルビア コルト ライフ スターレット

3 ブルース映画音楽 ポップス クラシック フラメンコ ラテン ゴスペル サンバ ラグタイム 校歌 ワルツ ハワイアン 童謡 ヒット曲 演歌 讃美歌 民謡 アニメソング バラード

のそれぞれマーチが持つ行進曲、自動車、スポーツカーに意味対応する共起語のセットが取れた。全て正しい語義のセットである。

b）「タブ」の例

1 :ツブラジイ ヤマモモ ユス クヌギ クスノキ ヒメユスリバ アラカシ ケヤキ シラカシ トベラ ミズキ ネズミモチ カシ ハチジョウススキ カゴノキ スダジイ ヤブニッケイ ヒバ

2 :ボタン マージン 検索窓 背景色 バックスペース 半角スペース 段落 インデント タイトル ウインドウ カンマ 大文字小文字変換 セミコロン プルダウンメニュー スクロールバー 段落書式 フォーム文字 禁則

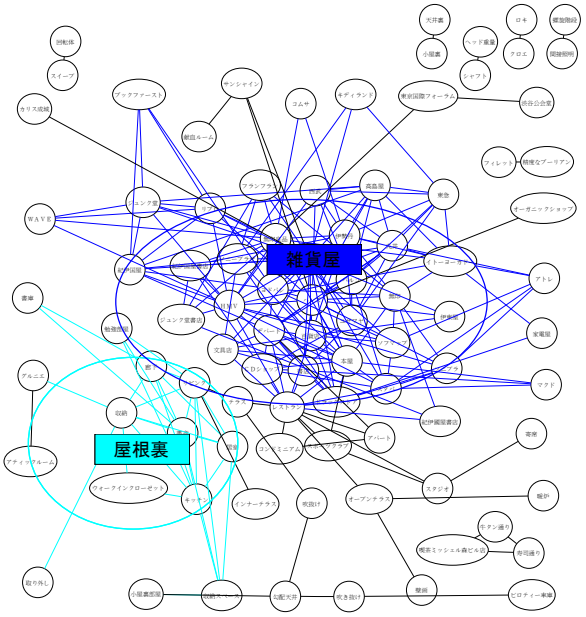

図 6　「ロフト」のクラスタリング結果 (Curvature を用いたクラスタリング)

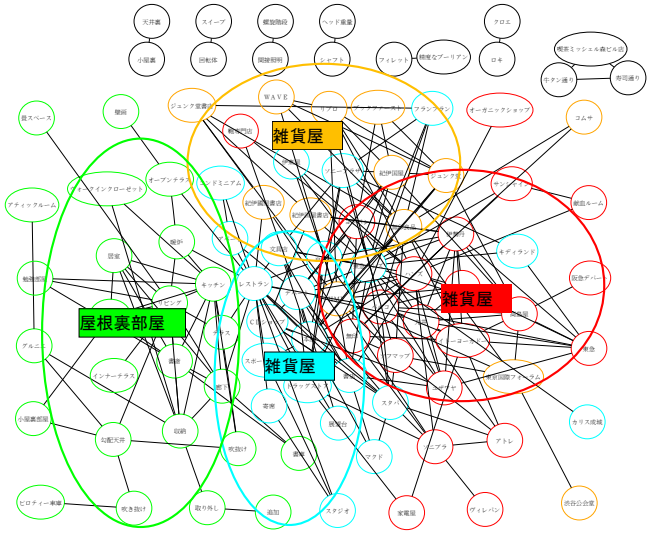

図 7　Newman 法を用いた「ロフト」のクラスタリング結果

処理 改行 ダイアログボックス タイトルバー ツールバー ツリー 記号 アドレスバー ウィンドウ スペース 角スペース メニュー リンク 改行文字スラッシュ

タブに関してはタブの木の意味と、タブキーに相当する意味に関する共起語集合が取れた。

c）「ロフト」の例

1: 百貨店 リブロ WAVE 文具店 丸井 ブックファースト 無印 ヨドバシ紀伊国屋書店 ジュンク堂書店 東京国際フォーラム

2:収納スペース キッチン 勉強部屋 書斎 廊下 居室 リビング

最後に共起語グラフを作りクラスタリングした結果の具体例について紹介する。「ロフト」をターゲットした場合、ベースラインである Curvture を用いた方法については図 6 のようにクラスタリングされた。Newman 法を用いた結果は図 7 のようにクラスタリングされ、3-クリークを用いたクラスタリング方法では図 8 のようにクラスタリングされた。

5. ま と め

この論文では情報検索における多義語が引き起こす問題について述べた。多義語が引き起こす問題は既存の曖昧性解消の技術を用いて解かれてきたが、既存の曖昧性解消の研究は多義語及び語義のセットが所与であることについて述べた。しかしこの問題設定であると、日々生まれる未知語や既知の多義語の語義の変遷を考慮しておらず、語義の自動抽出が重要であるということについて述べた。そして、語義の自動抽出について研究について大きく分けて 2 タイプ紹介した。既存の研究に対し

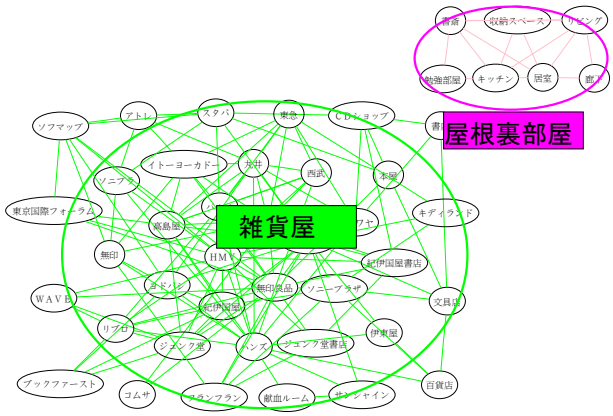

図 8　「ロフト」のクラスタリング結果 (3-クリークを用いたクラスタリング)

てグラフ構造を工夫し、単純な手法によっても一定の語義の自動抽出ができることを述べた。この方法の評価を行った。そしてその過程で wikipedia にも載っていな未知語義を獲得することもできた。例えばハクという語には魚のハクの意味、アニメの主人公ハクの意味があることが確認された。更に、本研究によって、多義語の各語義に対応した単語の集合が得られるが、この単語の集合は、多義語の曖昧性解消の教師データとして用いることができる。

文　献

[1] Beate Dorow, Dominic Widdows, and Katarina Ling. Using curvature and markov clustering in graphs for lexical acquisition.

[2] M.Newman and M.Girvan. Finding and evaluating community structure in networks. In *Physical Review E*, 2004.

[3] Patrick Pantel and Dekang Lin. Discovering word senses from text. In *In Proceedings of ACM SIGKDD*, pp. 613–619, 2002.