大規模コーパスからの語義のマイニング

Mining Word Sense from Large Corpora

田淵史郎 ♥ 鍛治伸裕 ♦
吉永直樹 ♠ 喜連川優 ♣

Shiro Tabuchi Nobuhiro Kaji
Noki Yoshinaga Masaru Kitsuregawa

自然言語処理における基礎技術の一つである語義曖昧性解消は，多義語の語義があらかじめ分かっているという問題設定でこれまで議論が行われてきた．しかし，ウェブのような大規模コーパスを処理対象とする場合，このような問題設定は必ずしも現実的ではない．そこで，ウェブのような大規模テキストにも適用可能な語義曖昧性解消の実現を目指し，多義語の語義をコーパスから自動的に発見する手法を提案する．語義の発見とは共起語のクラスタリングであるという考えに基づき，グラフ構造で表現された共起語集合にクラスタリングアルゴリズムを適用することによって，コーパスからの語義発見を試みた．74 の多義語に対して提案手法を適用したところ F 値で 0.77 という実験結果を得た．

Word Sense Disambiguation (WSD) is the problem that has long been investigated in NLP community. In conventional problem settings of WSD, it is assumed that all senses of polysemous words are identified in advance. This assumption, however, is obviously inappropriate in dealing with large Web corpora. To compensate for this situation, we present a method of discovering word senses from corpora.

1. はじめに

近年，ウェブに代表されるテキストデータは益々増加の一途を辿っており，そこからの価値創出，知識発見は重要な課題となっている．これに伴い，自然言語で記述されたテキストを計算機を用いて機械的に処理する必要性が高まりつつある．

テキストを計算処理する場合に発生する問題の一つが多義語の扱いである．自然言語では，一つの単語が文脈に応じて複数の異なる意味で使われることがある．例えば「ジャガー」という語は，車の意味で使われることもあれば，動物の意味として使われる場合もある．そのため，多義語がテキストに出現した場合には，それがどの意味で使われているのかを決定する必要がある．これは語義曖昧性解消と呼ばれ，自然言語処理における古典的な問題として知られている [9].

現在のところ，語義曖昧性解消は教師有り分類問題として解く方法が主流となっている．そのため，既存の方法で語義曖昧性解消を行おうとした場合，コーパス中に出現する多義語に対して，各語義に対応する訓練事例をあらかじめ作成しておく必要がある．しかし，ウェブのような大規模テキストを扱おうとする場合，そこにどのような多義語が出現しており，それがどのような語義を持っているのか必ずしも自明ではないため，訓練事例をあらかじめ作成しておくことが難しくなる．

本論文では，大規模なウェブコーパスにも適用可能な語義曖昧性解消の実現を目的として，多義語の語義をコーパスから自動的に発見する手法を提案する．語義の発見とは共起語のクラスタリングであるという考えに基づき，グラフ構造で表現された共起語集合にクラスタリングアルゴリズムを適用することによって，コーパスからの語義発見を試みた．これまでにも，共起語のグラフ構造にもとづいて語義発見を行う研究報告は存在するものの，その数は極めて少なく，また異なるクラスタリングアルゴリズム間での定量的比較もほとんど行われていない．そこで，我々はニューマン法，3-クリーク法というこれまでに試されていないアルゴリズムの有効性を検証するとともに，過去に提案されている Curvature 法との比較を行った．74 の多義語に対して 3 つの手法の比較実験を行ったところ，3-クリーク法が，F 値 0.77 という最も良い結果を得た．

2. 共起語クラスタリングに基づく語義発見

まず，何をもって多義語の「語義」を発見することができたと言えるか，ということを議論しておく必要がある．当然，これについては様々な立場がありうるが，本論文における基本的な考え方は「共起語のクラスタリングを語義の発見と捉える」というものである．以下，これについて詳しく説明をしていく．

多義語における語義と共起語の関係を見るため，例として「ジャガー」という多義語を考える．この単語は少なくとも車と動物の 2 つの語義を持っているが，もし「ジャガー」が前者の意味で使われた場合「ポルシェ」や「ベンツ」などの車に関連する語と共起することが予想される．一方，後者の意味であれば「ライオン」や「シマウマ」などが共起語になると考えられる．この例から分かるように，多義語の語義と共起語の間には対応関係を見て取ることができる．そこで我々は，語義の発見とは共起語をクラスタリングすることであると考えた．つまり，もし多義語の共起語を適切にクラスタリングすることが出来れば，得られたクラスタはその多義語が持つ語義に相当するであろうと考えた．

では，実際に共起語のクラスタを得ることが出来たとして，それがどのように語義曖昧性解消の高度化につながるのであろうか．共起語クラスタの利用方法として最初に考えられるのは，それを擬似的な訓練事例 [8] として用いることである．これによっ

♥ 野村総合研究所 s-tabuchi@nri.co.jp
♦ 東京大学 生産技術研究所 kaji@tkl.iis.u-tokyo.ac.jp
♠ 東京大学 生産技術研究所 ynaga@tkl.iis.u-tokyo.ac.jp
♣ 東京大学 生産技術研究所 kitsure@tkl.iis.u-tokyo.ac.jp

て，訓練事例を人手で作成するコストを大幅に削減することができる可能性がある．次に，得られた共起語クラスタを提示することによって，辞書や訓練事例の作成を支援することも考えられる．近年では，Wikipedia や WordNet をはじめとして，日本語処理のための言語資源の整備が進みつつある [4, 15]．しかしながら，ウェブのような大規模コーパスを処理対象に考えた場合，既存の言語自然には以前として量的，質的な問題が残っており，さらなる整備および拡張が重要な課題となっている．

3. コーパスからの語義発見手法

提案手法では，入力語の共起語をコーパスから取得してそれをグラフで表現する．このグラフでは共起語はノード，共起語間の類似性はエッジとして表現される．そして，グラフのノードをクラスタリングすることによって共起語のクラスタリング，すなわち入力語の語義を発見する．

3.1 共起語集合の取得

まず，入力語に対する共起語集合をコーパスから取得する．一般に，共起語の取得には，同一ウインドウに出現する語を取得する方法などが良く知られているが，ここではより正確な共起語を取得するために語彙統語パターンを用いた．以下の 4 つのパターンを形態素解析済みのテキストに適用して，X と Y にマッチした単語を共起語対として取り出す．このとき，X または Y のどちらか一方が入力語であった場合，残りの単語を入力語の共起語として収集する．

X や Y　X も Y も　X と Y と　X, Y,

3.2 グラフ表現

このようにして得られた共起語集合をグラフとして表現する．まず全ての共起語に対応するノードを作成する．そして，共起語間の意味的類似性を捉えるため，共起語同士が共起している場合にはノード間にエッジを作成する．共起語間の共起関係も，上と同じ語彙統語パターンを用いて求める．エッジには PMI の値を重みとして与える．2 つの共起語 w と w' の PMI は以下の式で求める．

$$\mathrm{PMI}(w, w') = \log \frac{f(w, w')f(*, *)}{f(w, *)f(*, w')}$$

ここで $f(w, w')$ は，2 つの共起語 w と w' が上記の語彙統語パターンにおいて共起した回数であり，$*$ は全ての単語について和を取ることを意味する．もし PMI の値が閾値 σ よりも小さい場合はそのエッジを除去する．以下では，こうして作成されたグラフのことを入力語の共起語グラフと呼ぶ．

3.3 グラフクラスタリング

共起語グラフに対してクラスタリングアルゴリズムを適用し，得られたクラスタから大きさが N 以下のものを除去する[1]．これによって，入力語の語義に特徴的な共起語クラスタを取得することができる．クラスタリング手法には以下の 2 つを用いた．

[1] 実験では $N = 3$ とした．

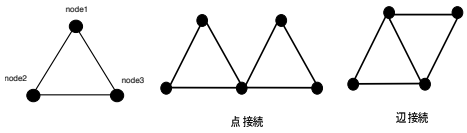

図 1　3-クリーク

3.3.1 Newman 法

まず一つめの手法として Newman 法を用いた [10]．Newman 法は，同一クラスタ (に含まれるノード) 間にはエッジが多く，異なるクラスタ間にはエッジが少なくなるようにノードをクラスタリングするアルゴリズムである．Newman 法では，クラスタリングの良さを以下の関数 Q で定義し，この値が大きくなるようにクラスタリングを行う．

$$Q = \sum_i \left(e_{ii} - a_i^2 \right) \tag{1}$$

ただし

$$e_{ij} = \frac{i \text{ 番目と } j \text{ 番目のクラスタ間のエッジ数}}{\text{全エッジ数}} \tag{2}$$

$$a_i = \sum_k e_{ik}. \tag{3}$$

である．

Newman 法では，通常のボトムアップクラスタリングと同様に，はじめにノードと同じ数だけをクラスタを作成する．そして，Q の値が最も大きく増加するように，2 つのクラスタを統合する．i 番目と j 番目のクラスタを統合したときの Q の増分は以下の式で求めることができる．

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j). \tag{4}$$

上記の処理は Q の値が極大値に達するまで続ける．

語義発見における Newman 法の利点の一つは，クラスタ数 (= 語義数) をあらかじめ設定しておく必要がないことである．もし K-means のようなアルゴリズムで語義発見を行う場合には，全ての入力語についてあらかじめ適切なクラスタ数を設定しておく必要がある．一方，Newman 法においては，Q の極大解が発見された時点でアルゴリズムが終了し，それに伴いクラスタ数も自然に決定される．

3.3.2 3-クリーク法

2 つ目の手法として，3-クリークに着目したシンプルなボトムアップクラスタリングアルゴリズムを考案した．この手法では，図 1 左側のような 3 つのノードで構成されるクリーク (3-クリーク) に着目する．共起語グラフにおいて 3-クリークを構成する共起語は，互いに類似した意味を持っていると考えられる．そこで，共起語グラフ内の 3-クリークを最小単位とし，点接続または辺接続の関係にある 3-クリーク (図 1 右側) を順次統合していくことによって，共起語のクラスタリングを行う (Algorithm 1).

Algorithm 1 3-クリークに基づく共起語クラスタリング

入力: 共起語グラフ $G = (V, E)$
出力: 共起語クラスタ集合 $\mathcal{C}$

```
C ← φ
while V ≠ φ do
    C_v ← pop(V)
    while ∃v ∈ C_v, ∃(u, u') ∈ E s.t. (v, u), (v, u') ∈ E do
        C_v ← C_v ∪ {u, u'}
        E ← E − {(u, u')}
    end while
    C ← C ∪ C_v
    V ← V − C_v
end while
return C
```

4. 実験

本節では，実際に多義語に対し Curvature 法，Newman 法，および 3-クリーク法を適用し，提案手法である 3-クリーク法の有効性を確認する．

4.1 データセット

まずテストデータとして，Wikipedia[2]の曖昧さ回避のためのページなどを参考に，人手で多義語を 74 語用意した．次に，Web から自動収集した約 1 億 7000 万文から，前節に述べた手法により各多義語の共起語を収集し，それぞれ共起語グラフを構成した．このようにして得られた共起語グラフを入力として，Newman 法と 3-クリーク法により共起語のクラスタリングを行い，各多義語に対する共起語のクラスタを得た．なお，比較のために Curvature 法 [2] を用いた同様の実験も行った．Curvature 法の詳細は文献を参照されたい．

4.2 評価方法

まず，各多義語に対して得られた共起語クラスタに対し，以下の手順に従い，人手で正解／不正解のラベルを付与した．

1. 多義語の共起語集合を参考にして，語義を人手で列挙する．
2. 得られた各共起語クラスタが，1 で列挙した語義に対応するとき，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するとき（過分割）には，一つのクラスタのみ正解とし，残りは不正解とする．

このようにして各クラスタに付与された正解ラベルをもとに，以下の尺度を用いて，各多義語に関する共起語クラスタの評価を行った．

$$\text{再現率} = \frac{\text{正解ラベルを付与したクラスタの数}}{\text{列挙した語義の数}} \quad (5)$$

$$\text{適合率} = \frac{\text{正解ラベルを付与したクラスタの数}}{\text{全クラスタの数}} \quad (6)$$

$$F\ \text{値} = \frac{2 \cdot \text{再現率} \cdot \text{適合率}}{\text{再現率} + \text{適合率}} \quad (7)$$

[2] http://ja.wikipedia.jp/

表 1 多義語「ロス」の共起語のクラスタリング結果に対する語義判定

ID	クラスタに含まれる共起語	語義判定
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	×（1 と重複）
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	×

表 2 共起語クラスタリングの実験結果

手法	適合率	再現率	F 値	平均語義数
Curvature	0.63	0.66	0.65	1.6
Newman	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

例えば，「ロス」という多義語に対して，表 1 のような共起語クラスタが得られた場合，その適合率，再現率は以下のように計算される．まず「ロス」という語の共起語集合から人手で列挙された語義は，地名のロサンゼルスと損失を意味するロスの二つであった．この語義をもとに表 1 のクラスタを順に確認すると，クラスタ 1 は地名の意味に対応し，正解となる．次に，クラスタ 2 も地名の意味に対応しているが，既にクラスタ 1 に地名の語義として正解ラベルを与えているので，このクラスタは不正解となる．クラスタ 3 は損失の意味に対応し，正解となる．クラスタ 4, 5 は語義に対応しないので，共に不正解のラベルを付与する．これらの正解／不正解ラベルから，表 1 の「ロス」に関する再現率は $2/2 = 1.0$, 適合率は $2/5 = 0.4$ と計算される．F 値は式 7 から 0.57 となる．

4.3 実験結果

Web から抽出した多義語 74 語に対する Curvature 法，Newman 法，3-クリーク法による共起語クラスタリング結果を表 2 に示す．手法の再現率，適合率，F 値は，各多義語に対して得られた共起語クラスタの再現率，適合率，F 値のマクロ平均を用いた．平均語義数は，各手法により発見された語義（共起語クラスタ）の数である．Curvature 法では，curvature の値が 0.35 以下のノードを削除し，その上で繋がったノードを一つのクラスタとした．3-クリーク法では，PMI に基づくエッジカットを行い，閾値 $\sigma = -\infty, 0, 5$ のときの結果を調べた．

実験結果から，適合率，F 値に関して我々の提案する 3-クリーク法 ($\sigma = 5$) により最も良い結果が得られることが分かった．特に，PMI に基づくエッジカットの閾値を上げることで，適合率，再現率共に大幅に改善できることを確認した．また，再現率については，Newman 法が最も良い結果となった．テストデータに

表 3　3-クリーク法により獲得された語義クラスタ

多義語	獲得した語義クラスタ
マーチ	リングマーチトライアンフ ＢＳＡ ミニ
	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
	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
タブ	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
	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
ロフト	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
	収納スペース キッチン 勉強部屋 書斎 廊下 居室 リビング

含まれる多義語に対し，人手で列挙された平均語義数は，2.3 個であり，最大語義数は 4 個であった．

次に，多義語「マーチ」,「タブ」,「ロフト」にに対して 3-クリーク法で得られた語義クラスタを表 3 に示す.「マーチ」ではそれぞれスポーツカー，自動車，行進曲の意味の共起語クラスタが得られている．共起語クラスタとして語義を表現することで，概念的に上位下位の関係にあり，本来区別の難しいスポーツカーと自動車の語義が区別されていることは興味深い．一方,「タブ」については,「（植物の）タブノキ」,「Tab キー」に相当する共起語クラスタが,「ロフト」については「（雑貨屋）の LOFT」「屋根裏部屋」に相当する共起語クラスタが得られた．どちらの多義語についても，共起語の集合として語義を捉えることで，明確に異なる複数の語義を持つことが確認できる．

最後に，多義語「ロフト」の共起語グラフに関する，Curvature 法，Newman 法，3-クリーク法によるクラスタリング結果を図 2, 3, 4 にそれぞれ示す．どの例でも「ロフトの」の語義「屋根裏部屋」、「雑貨屋」が得られているもの，Newman 法では同一の語義に対して複数のクラスタが得られてしまっていること

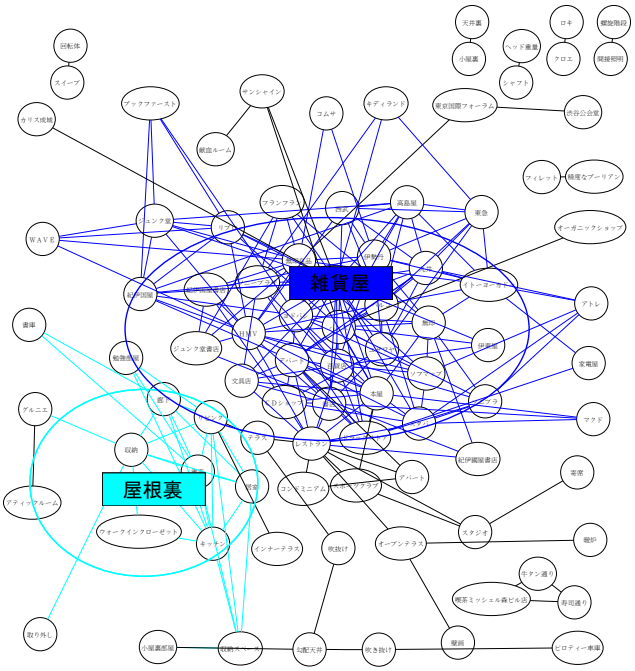

図 2　Curvature 法による多義語「ロフト」のクラスタリング結果

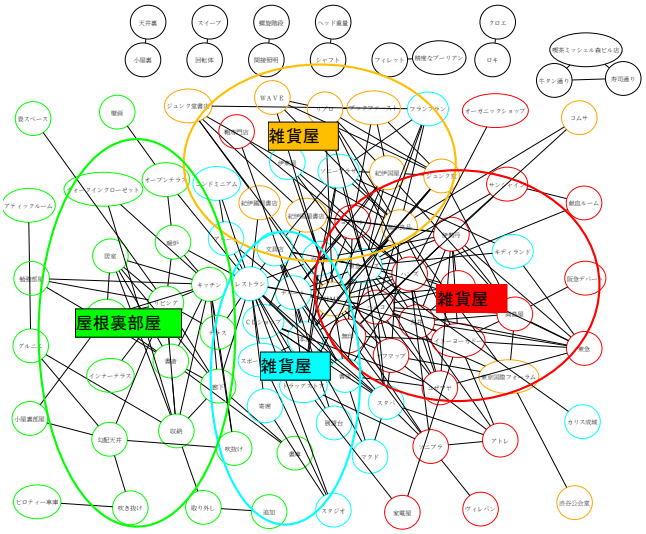

図 3　Newman 法による多義語「ロフト」のクラスタリング結果

（過分割）が分かる．

5.　関連研究

共起語をグラフで表現し，それをクラスタリングすることによって語義発見を行ったのは本論文が初めてではない [1, 2, 14]. しかし，これまで語義発見に関する研究例は極めて少なく，異なるアルゴリズム間の定量的な比較はほとんど行われていない．そ

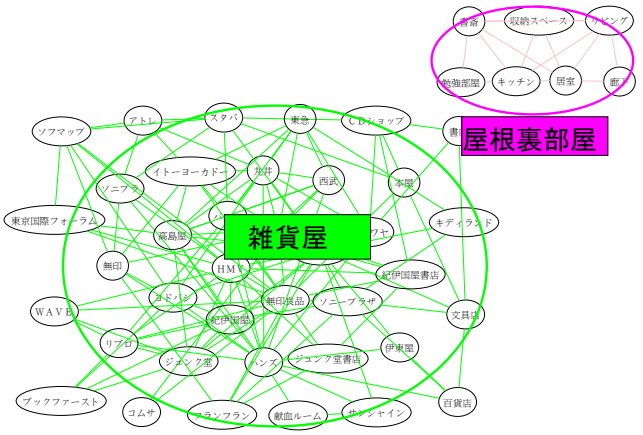

図 4　3-クリーク 法による多義語「ロフト」のクラスタリング結果

のため，どのようなクラスタリングアルゴリズムが有効であるかは十分に議論がつくされていなかった．これに対して，我々はニューマン法，3-クリーク法というこれまでに試されていないアルゴリズムの有効性を検証するとともに，過去に提案されている Curvature 法との比較を行った．今後は，混合モデルに基づくクラスタリングアルゴリズムなどの有効性の検証も行っていきたい [11].

Lin と Pantel は，語義発見のためのクラスタリングアルゴリズム Clustering By Committee (CBC) を考案した [7, 12]．彼らのクラスタリング手法はグラフに基づくものではなく，分布類似度 [6] に基づくものとなっており，この点において提案手法と異なる．グラフに基づく手法と分布類似度に基づく手法は，互いに相補的な性質を持っているため [5]，今後はこの 2 つの手法を統合したアルゴリズムを検討していきたいと考えている．

Erk は外れ値検出に基づく新語義発見手法を考案した [3]．彼が議論しているのは辞書などの外部知識に記述されていない新しい語義を検出する問題であり，本研究とは問題設定が異なるものの関連が深い．語義発見において外部知識をどうに活用すべきかということは重要な問題であるが，ウェブに出現する単語には，そもそも辞書などの外部知識に登録されていない未知語も多い．そのため，ウェブテキストを処理する場合には，提案手法のような外部知識にできる限り依存しない方法が重要になると考えている．

6. おわりに

本論文では大規模なコーパスから多義語の語義を発見するための手法について述べた．語義発見を共起語のクラスタリング問題と考え，2 つのクラスタリングアルゴリズム (Newman 法と 3-クリーク法) の適用を試み，それぞれの有効性の検証を行った．実験の結果，我々の提案する 3-クリーク法が良い結果が得られることが分かった．

実験では，得られた共起語クラスタの妥当性は人手によって判定した．これを語義曖昧性解消のための擬似訓練事例として用いてその性能を検証するなど，得られたクラスタの有用性に対するより詳細な分析は今後の課題である．また，現在のシステムの出力は単語クラスタのみであり，どのクラスタがどの語義に対応しているのか直感的に理解しづらくなっているため，クラスタへのラベル付けにも今後取り組む予定である．

[文献]

[1] Dorow, B. and Widdows, D.: “Discovering Corpus-Specific Word Senses”, In Proc. of EACL, pp. 79-82, (2003).

[2] Dorow, B., Widdows, D., Ling, K., Echmann, J., Sergi, D., and Moses, E.: “Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination”, In Proc. of MEANING, (2005).

[3] Erk, K.: “Unknown Word Sense Detection as Outlier Detection”, In Proc. of NAACL, pp. 128-135, (2006).

[4] Isahara, H., Bond, Uchimoto, K., Utiyama, M., and F., Kanzaki, K.: “Development of Japanese WordNet”, In Proc. of LREC, (2008).

[5] Kaji, N., and Kitsuregaw, M.: “Using Hidden Markov Random Fields to Combine Distributional and Pattern-based Word Clustering”, In Proc. of COLING, pp. 401-408, (2008).

[6] Lin, D.: “Automatic Retrieval and Clustering of Similar Words”, In Proc. of COLING/ACL, pp. 768-774, (1998).

[7] Lin, D. and Pantel, P.: “Concept Discovery from Text”, In Proc. of COLING, pp. 577-583, (2002).

[8] Lu, Z., Wang, H., Yao, J., Liu, T., and Li, S.: “An Equivalent Pseudoword Solution to Chinese Word Sense Disambiguation ”, In Proc. of COLING-ACL, pp. 457-464, (2006).

[9] Manning, D., C., and Schütze, H.: “Foundations of Statistical Natural Language Processing”, MIT Press, (1998).

[10] Newman, M. and Girvan, G.: “Finding and Evaluating Community Structure in Networks”, In Physical Revew, E, (2004).

[11] Newman, M. and Leicht, E.: “Mixture Models and Exploratory Analysis in Networks”, In Proc. of National Academy of Sciences **104** (23), pp. 9564-9569 (2007).

[12] Pantel, P. and Lin, D.: “Discovering Word Senses from Text”, In Proc. of KDD, pp. 613-619, (2002).

[13] Purandare, A. and Pedersen, T.: “Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces”, In Proc. of CoNLL, pp. 41-48, (2004).

[14] Widdows, D. and Dorow, B.: “A Graph Model for Unsupervised Lexical Acquisition”, In Proc. of COLING,

(2000).
[15] 白井清昭: “SENSEVAL-2 日本語辞書タスク”, 自然言語処理 **10**(3), pp. 3-24, (2003).

田淵史郎 Shiro Tabuchi
2009 年東京大学大学院情報理工学系研究科修士課程修了．同年より野村総合研究所勤務.
鍛治伸裕 Nobuhiro Kaji
2005 年東京大学大学院情報理工学系研究科博士課程修了．情報理工学博士．現在，東京大学生産技術研究所特任助教．自然言語処理の研究に従事.
吉永直樹 Noki Yoshinaga
2005 年東京大学大学院情報理工学系研究科博士課程修了．2002 年より 2008 年まで日本学術振興会特別研究員（DC1，PD）．2008 年 4 月より東京大学生産技術研究所特任助教．博士 (情報理工学)．自然言語処理の研究に従事.
喜連川優 Masaru Kitsuregawa
1978 年東京大学工学部電子工学科卒業．1983 年同大大学院工学系研究科情報工学専攻博士課程修了．工学博士．同年同大生産技術研究所講師．現在，同教授．2003 年より同所戦略情報融合国際研究センター長．データベース工学，並列処理，Web マイニングに関する研究に従事．現在，日本データベース学会理事，情報処理学会，電子情報通信学会各フェロー．ACM SIGMOD Japan Chapter Chair，電子情報通信学会データ工学研究専門委員会委員長歴任．VLDB Trustee(1997-2002), IEEE ICDE, PAKDD,WAIM などステアリング委員．IEEE データ工学国際会議 Program Co-chair(99),General Co-chair(05).