Vol.2010-NL-199 No.14
2010/11/19

ハブを作らないグラフ構築法を用いた半教師あり語義曖昧性解消

小嵜耕平[†1]　新保仁[†1]
小町守[†1]　松本裕治[†1]

グラフに基づく半教師あり分類問題に対する手法において，グラフ構築の方法が結果に大きな影響を与えることが報告されており，近年関心を集めている．本稿ではグラフのハブが与える分類器への影響に注目し，効率的な計算が可能な新しいグラフ構築法を提案する．また語義曖昧性解消タスクにおいて，提案するグラフ構築法を用いることで，既存のグラフ構築法を用いた場合と比較して，高い分類性能が得られることを示す．さらに提案するグラフ構築法は他の既存のグラフ構築法と比較し，推論アルゴリズムの根拠となるクラスタ仮説をよりよく満たしていることを確認した．

Semi-supervised Word Sense Disambiguation using Degree Bounded Graph Construction

Kohei Ozaki,[†1] Masashi Shimbo,[†1]
Mamoru Komachi[†1] and Yuji Matsumoto[†1]

In graph-based semi-supervised learning, graph construction methods are known to influence performance and have gained increasing attention in research communities over the past few years. In this papers, we investigate the effect of the hub nodes of a graph on a classification task and propose a new efficient graph construction method. In word sense disambiguation task, we demonstrate our proposed method outperforms the current state-of-the-art graph construction method in prediction accuracy. In addition, we show the graph produced by our proposed method satisfies cluster assumption better than those by other methods.

1. はじめに

多義語に対して周辺の文脈情報を用いることで，語義に基づく分類を行う語義曖昧性解消タスクは，自然言語処理における中心的なタスクの一つであり，統計的機械翻訳などの応用分野においてその有効性が示されている．[1),2)] 語義曖昧性解消タスクに対して，いままでに様々なアプローチが提案されてきた．教師あり学習を用いた語義曖昧性解消では，高い分類精度を達成しているものの，そのためには大量の人手により整備された言語資源が必要となる．語義曖昧性解消タスクは教師データを用意するために必要なアノテーションのコストが高いため，教師あり学習にとっての十分な教師データが存在しない場合が多い．[3)] 一方で，半教師あり学習のアプローチを用いると，少量の教師データと大量のラベルなしデータを用いることで，高い分類性能を得ることができる．本稿では，人手による手間を最小限に抑える一方で，高い分類性能を持つ半教師あり学習による語義曖昧性解消システムの開発を目標とする．

半教師あり学習は近年盛んに研究されている分野である．中でもデータ間の類似関係により定義されるグラフに基づく半教師あり学習のアプローチは，近年ますます注目を浴びている．[4)–6)] 自然言語処理における様々なタスクにおいても，グラフに基づくアプローチが有効であるという報告がなされている．[7)–10)] グラフに基づく半教師あり学習では，分類精度の向上と計算の効率化を目的として，グラフ構築のプロセスにおいて辺をカットするグラフスパース化がしばしば行なわれており，グラフ構築の方法が分類やクラスタリングの結果に大きな影響を与えることが報告されている．[11)–14)] グラフスパース化には一般的に k-近傍グラフが性能の保証なく，しばしば用いられてきたが，最近の研究では b-マッチンググラフと呼ばれるグラフを用いることが提案されており，半教師あり分類問題やクラスタリングなどにおいてその有効性が示されている．[12),15)]

半教師あり語義曖昧性解消タスクや訳語曖昧性解消タスクにおいても，いままでに様々な研究がされてきた．Yarowsky の Bootstrapping を用いた研究[16)]，Pham らの Spectral Graph Transducer と co-training を用いた研究[8)]，Li & Li の Bilingual Bootstrapping を用いた研究[17)]，Niu らのラベル伝播法を用いた研究[7)] などがあり，中でも Niu らは k-近傍グラフを用いたラベル伝播法を用いることで，Bootstrapping による手法や SVM を用い

†1 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

た教師あり学習の手法[18] と比較して，より優れた分類性能を達成することを示した．

小嵜らは Niu らの研究に対して，ラベル伝播法などのグラフを用いた推論アルゴリズムにおいてはグラフの構築法が重要であることを示した最近の研究[12),15)] に着目し，Niu らが用いた k-近傍グラフではなく b-マッチンググラフを用いることで，分類精度の改善が見込めることを確認した．[13)]

本稿では更に分類性能を改善し，なおかつ b-マッチンググラフでの問題点であった効率的な計算を可能にする新しいグラフ構築法を提案する．b-マッチンググラフはグラフ構築に必要な計算量が大きい一方で，k-近傍グラフの構築については様々な近似計算の研究がなされており，[19)–22)] 提案するグラフ構築法も k-近傍グラフにおける近似計算の手法が利用できるため，b-マッチンググラフでは困難であった大規模データへの適用の実現も期待できる．

本稿の構成は以下の通りである．2 節にて語義曖昧性解消タスクを半教師あり学習の文脈で定式化し，グラフに基づく半教師あり学習のアルゴリズムについて記述する．3 節では本研究の貢献にあたるグラフ構築における提案手法について述べ，4 節で語義曖昧性解消タスクにおける評価データセットを用いた実験結果を与え，グラフ構築法の比較をする．最後に我々の今回の仕事をまとめ，今後の課題について述べる．

2. グラフに基づく半教師あり語義曖昧性解消

2.1 語義曖昧性解消タスクの定式化

本節では，語義曖昧性解消タスクを半教師ありの分類問題に定式化する方法について述べる．基本的な記号などは小嵜らの論文[13)] と同様の定義を用いる．

語義に曖昧性のある語を w とし，この語が現われる文脈の集合を $X = \{x_i\}_{i=1}^{n}$ とする．x_i は i 番目に w が現われる文脈であり，n は文脈の総数である．$S = \{s_j\}_{j=1}^{c}$ は曖昧性のある語 w における，割り当てが可能なクラスラベル（語義）の集合である．はじめの l 個の文脈 x_g $(1 \leq g \leq l)$ をラベルありデータとし，他の u $(l+u=n)$ 個の文脈 x_h $(l+1 \leq h \leq n)$ をラベルなしデータとする．半教師あり分類問題の目的は，文脈 x_h に現れる語 w の語義ラベル $(\in S)$ を，x_g のラベル情報と文脈集合 X の類似情報を使うことで推定することである．

文脈集合 X の任意の 2 文脈間 x_i, x_j 間は，何らかの類似度尺度を与えることで，2 文脈間がより似ているなら大きな重みを与える．ここで，文脈集合 X に一対一対応する頂点集合を V とし，文脈間類似度を辺の重みとする辺集合を E，辺の重みを要素として持つ $n \times n$ 行列を W として定義することで，重み付きグラフ $G = (V, E, W)$ を定める．グラフを用いる半教師あり語義曖昧性解消では，この重み付きグラフ G を入力として，x_h $(l+1 \leq h \leq n)$ に対応するクラスラベルを当てる問題として定式化できる．

2.2 グラフに基づく半教師あり学習法

データ間の類似度によるグラフ G を用いて分類問題を解くアプローチは，グラフスパース化によるグラフの構築と，推論アルゴリズムの適用の 2 つのステップに分けることができる．これら 2 つのステップについてそれぞれ述べる．

2.3 グラフ構築

重み付き完全グラフ，あるいはその部分グラフ G から，よりスパースな重み付き部分グラフを見つけるグラフスパース化は，分類精度や計算効率に大きく影響する重要な手続きである．推論アルゴリズムが辺の数やスパース化された頂点の次数に従う計算量により定義されている場合は，グラフスパース化は計算の効率化に繋る．また，データ間の類似度尺度によっては，近いデータにのみ類似度が有用であり，相対的に遠く離れたデータ間の類似度は信頼できない場合があり，グラフスパース化によりそれらの信頼度の低い類似度を無視することで，分類精度の向上も期待できる．

グラフスパース化の手続は，グラフ G の辺をカットすることにより行われる．任意の i, j 成分に対して，1 ならば $(i,j) \in E$ の辺をカットせず，0 ならば $(i,j) \in E$ をカットすることを表現する行列として，$n \times n$ の二値行列 P を導入し，グラフスパース化を行列 P を求める問題を考える．

2.3.1 k-近傍グラフ

k-近傍グラフはグラフスパース化の方法として広く一般的に使われてきた．[7),9),23)] ある閾値 ε 以下の重みを持つ辺をカットする ε-近傍グラフが用いられることもあるが，この方法は孤立点を作りやすくパラメータの設定が困難な方法であるため，k-近傍グラフと比べると低いパフォーマンスを得る傾向にある．[11)]

k-近傍グラフは以下の最適化で得られるグラフとして定式化できる．

$$\max_{\hat{P}} \sum_{i,j} \hat{P}_{ij} W_{ij} \quad s.t. \sum_{j} \hat{P}_{ij} = k,\ \hat{P}_{ii} = 0,\ \forall i, j \in 1, \ldots, n \tag{1}$$

そして，すべての i, j に対して $P_{ij} = \max(\hat{P}_{ij}, \hat{P}_{ji})$ を求めることにより，二値行列 P を得る．この貪欲なアルゴリズムでは，得られるグラフの次数が $\sum_j P_{ij} \geq k$ と均一でないことから，クラスタ間を結ぶ辺が作られやすいと考えられる．

2.3.2 b-マッチンググラフ

そこで，k-近傍グラフに代る手法として，より正則なグラフに近い b-マッチンググラフ

Vol.2010-NL-199 No.14
2010/11/19

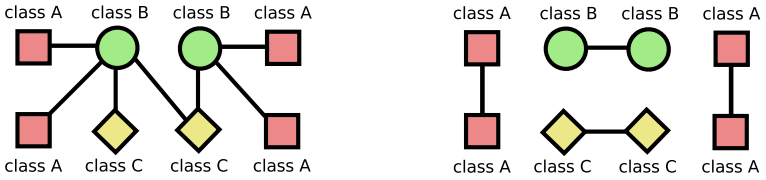

図 1 グラフに含まれる ϕ-edge の割合が大きい場合と小さい場合の例

を分類問題に応用する提案がなされている.[15] b-マッチンググラフは以下の最適化を行なうことによって得られるグラフである.

$$\max_{P} \sum_{i,j} P_{ij} W_{ij} \quad s.t. \sum_{j} P_{ij} = k,\ P_{ii} = 0,\ P_{ij} = P_{ji},\ \forall i, j \in 1, \ldots, n \tag{2}$$

この最適化により得られる行列 P は，対称性の制約を持つため，得られるグラフの次数が均一になる特徴がある.

2.4 ラベル伝播法による語義ラベルの推定

グラフに基づく半教師あり学習のアプローチでは，構築されたグラフを用いてクラスラベルを推定する．大抵の手法はクラスラベルを近隣頂点に伝播する手法である．後の実験では，Zhu らのラベル伝播法[4] を用いた．この手法に関する説明は付録 A.1 にて記述する.

3. ハブに着目したグラフ構築法

グラフに基づく半教師あり学習に対して，k-近傍グラフではなく，最近 b-マッチンググラフと呼ばれる頂点次数を均一にするグラフを用いることが提案されており，小嵜らはその効果を語義曖昧性解消タスクにおいて確認しているが，このグラフを用いることで分類精度が向上したのはいわゆるハブが結果として作られなくなったからではないか，との仮説が立てられる．以下ではこの仮説を検証し，その結果に基づいて新しい，単純ではあるが効果的なグラフ構築法を提案する.

3.1 グラフに基づく手法におけるハブの影響

グラフにおける頂点次数の大きな頂点をハブと呼ぶことにする．ハブを取り除くことが分類性能の向上に繋がることを確認するために予備実験を行った.

予備実験のためにグラフの辺に対して, Cesa-Bianchi らにより導入された ϕ-edge の定義を用いる．ϕ-edge とは，頂点にクラスラベルが付与されているグラフについて，辺をなす 2 頂点に付与されたクラスラベルが異なる辺のことをいう.[24] 例えば，**図 1** のようなクラスラベルが与えられたグラフにおいて，左のグラフはすべての辺が ϕ-edge で構成されたグラフとなり，右のグラフはすべての辺が ϕ-edge でない辺で構成されたグラフとなる．近隣頂点にクラスラベルを伝播する手法を用いるために，ϕ-edge をできるだけ取り除くことが望ましいが，本研究における問題設定では，大抵の頂点にはクラスラベルが付与されていないので，ふつう辺が ϕ-edge かどうかはわからない.

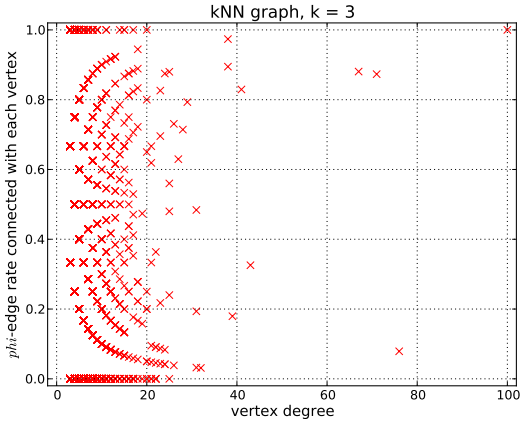

図 2 各頂点の頂点次数と頂点に接続する ϕ-edge の割合

表 1 k-近傍グラフと，k-近傍グラフからハブに接続する辺を削除したグラフとの平均分類精度の比較

	Averaged accuracy	
d	original	eliminated
0	52.6	—
1	65.9	**66.6**
2	65.7	**66.0**
3	69.7	69.7
4	100.0	100.0
—	65.9	**66.2**

予備実験では，後述する語義曖昧性解消タスクにおける line データセット[25] に対して，ランダムに 10 通りの 10%のサンプルをラベルありデータとして作り，k-近傍グラフと，k-近傍グラフからハブに接続する辺を削除したグラフを用いてラベル伝播法[4] を適用し，平均精度を比較することでハブを取り除く効果を確認する．k-近傍グラフのパラメタ k は 3 と固定し，k-近傍グラフと，k-近傍グラフにおける頂点次数が 30 以上の頂点をハブとして接続する辺を取り除いたグラフとを比較した.

図 2 は k-近傍グラフの各頂点について，すべてのクラスラベルが与えられていたとき，頂点次数と頂点に接続する辺のうち ϕ-edge である割合を調べたものである．この図から，k-近傍グラフには頂点次数が大きく頂点に接続する ϕ-edge の割合の高いハブができていることが確認できる.

表 1 では k-近傍グラフにおいて，取り除くハブからの最短パスの長さを d とし，d に応じて頂点に対応するサンプルの分類精度に変化が見られるかを確認した．ただし $d = 0$ は

ハブ自身を表す．予備実験において，k-近傍グラフを用いた場合のハブに対応するサンプルの分類結果はラベル伝播法の分類結果をそのまま用い，ハブを取り除いたグラフにおいては取り除いたハブに対応するサンプルをすべて負例として扱った．表から，k-近傍グラフからハブを取り除くことで，ハブの近くにあった頂点の分類精度が向上していることが確認できる．また，全体の平均分類精度においてもハブを取り除いたグラフを用いた方法がハブに対応するサンプルをすべて負例として扱う不利な設定であったにも関らず，より良い分類精度を達成している．

3.2 頂点次数に上限の制約を与えた提案グラフ

本稿では，任意の 2 頂点の互いの頂点が，互いの k 近傍に含まれている場合にのみ辺を作ることで得られるグラフを，グラフに基づく半教師あり学習に用いることを提案する．以下，このグラフを**次数制約付きグラフ**と呼ぶ．このグラフは 式 (1) の行列 $\hat{P}$ を用いて，すべての i, j に対して $P_{ij} = \min(\hat{P}_{ij}, \hat{P}_{ji})$ を求めることで得られる二値行列 P として表現できる．次数制約付きグラフは V, W を入力としたとき，以下の手順で構築できる．

(1) 各頂点に応じてヒープを $|V|$ 個用意する．
(2) 各頂点 v について，類似度を重みとして持つ，v に接続する辺をすべてヒープに挿入する．
(3) 各頂点に対応するヒープから類似度の大きいものを k 個ずつ取り出したとき，重複して現われる辺集合を E とする．

この手続きにより次数制約付きグラフ $G = (V, E, W)$ が得られる．

このグラフは結果的にパラメタ k 以下に頂点次数が制限され，ハブを作らないグラフが，k-近傍グラフの部分グラフとして構築される．

3.3 グラフ構築の計算量

k-近傍グラフはフィボナッチヒープを用いることで $O(n^2)$ でのグラフ構築が可能であり，また様々な近似計算法が提案されている．[19)–22)]．b-マッチンググラフは NP-困難な組み合わせ最適化問題を解くことにより得られるグラフであり，この最適化はループあり信念伝播法を用いることで多項式時間での近似解を求めることができる[15),26)] が，$O(n^3)$ の計算量が必要であるため，比較的計算量が高く使いづらい．その一方で，次数制約付きグラフは k-近傍グラフと同様にフィボナッチヒープを用いることで 3.2 節の手順により $O(n^2)$ でのグラフ構築が可能である．また，次数制約付きグラフは k-近傍グラフの部分グラフでもあることから，k-近傍グラフを構築する際にできる行列 $\hat{P}$（あるいはヒープ）を用いることで，k-近傍グラフから $O(kn)$ の計算量でグラフを構築することができ，k-近傍グラフのための高速な近似計算法を用いることができる．

さらに，k-近傍グラフや提案するグラフ構築法においては，パラメタ k のグラフ構築の過程で使用するフィボナッチヒープを保持しておくことで，$O(kn)$ の計算量でパラメタ $k+1$ のグラフが構築可能である．一方で b-マッチンググラフでは，パラメタ b でのグラフ構築の計算結果を使いパラメタ $b+1$ のグラフ構築の計算を効率化することができない．

3.4 グラフを連結するヒューリスティック

上述のグラフ構築法はグラフの連結性を保証するものではない．連結成分数が多くなるに従い，同じ連結成分中にラベルありサンプルがない場合，その連結成分に含まれるラベルなしサンプルは，ラベル伝播法など推論アルゴリズムによってはラベルを推定することができない．次数制約付きグラフは連結成分数を多く作る傾向にあるため，この問題に直面する．このようなラベルの推定ができなくなることを避けるために，グラフの連結成分数を 1 つにする必要がある．本稿ではデータセット全体の最大全域木の辺を加えるヒューリスティックを提案手法に対して適用した．

4. 関 連 研 究

Radovanović らによると，サンプルが高次元データである場合には，他のサンプルの k-近傍に含まれることがよく起きるハブサンプル[⋆1]が作られる傾向があり，このとき各サンプルにクラスラベルが付与されているとすると，k-近傍にハブサンプルを含むサンプルと，ハブサンプルとクラスラベルが異なることが多い「悪いハブサンプル」が k-近傍分類器に悪い影響を与えることが報告されている．[27)] この研究から，似たクラスラベルを持つものは似たデータであるというクラスタ仮説を根拠としたグラフに基づく半教師あり学習の手法においても，k-近傍グラフにおいて，接続する辺が ϕ-edge である割合の大きなハブノードが，分類性能に悪い影響を与えることが示唆される．

5. 語義曖昧性解消タスクでの評価

5.1 評価データセット

語義曖昧性解消タスクにおけるグラフ構築法の比較のために，我々は英語の名詞における語義曖昧性解消タスクに広く用いられている Pedersen の評価データセット[25)⋆2] を用いて

⋆1 Radovanović らの論文ではこのサンプルをハブと呼んでいるが，ここではグラフのハブと区別するために，便宜上ハブサンプルと呼ぶことにする．
⋆2 http://www.d.umn.edu/˜tpederse/data.html

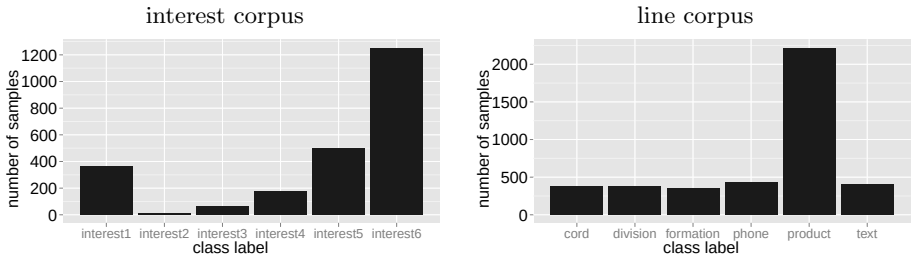

図 3 2 つの名詞 interest,line の語義ラベルの分布

実験を行った．このデータセットは，2 つの名詞 interest，line に対して，主に Wall Streat Journal の記事から名詞の interest，line が用いられる文脈を取り出し，タグ付けをしたもので構成される． さらに Senseval-3 English Lexical Sample (S3LS)[18)⋆1] タスク内の全名詞に対しても実験を行った．

教師データが少量しか得られない状況を人工的に作るために，訓練データとテストデータを合わせた全体に対して，ランダムに 10 通りの 10%のデータをラベルありデータとして，残りをテストデータとして用いる．また Niu らの研究と同様に，サンプルに複数のクラスラベルがある場合には，クラスラベルを一つのみとしてデータセットを扱う．それぞれの語は複数のクラスラベルの候補を持ち，各クラスラベルの出現頻度には大きな偏りがある (**図 3**)．語義曖昧性解消タスクにおいては，各語義の出現頻度の偏りはよく現われるので，最頻出語義を常に分類結果として答える分類器がベースラインとしてしばしば設けられる．

5.2 実験結果

特徴量の取り方と類似度尺度は Niu らによる先行研究と同様のものを用いた．特徴量は，曖昧性を解消したい語の周辺の語の品詞とその語の位置，順序を考慮しない周辺文脈に出現する語，そして局所的なコロケーションを用いた．[28)] また類似度はコサイン尺度を用いた．

Pedersen の interest, line データセットを用いて，グラフ構築法における k-近傍グラフ (kNN)，b-マッチンググラフ (bM)，次数に上限の制約を与える次数制約付きグラフ (degree bounded; DB)，そして k-近傍グラフから頂点次数 d 以上の頂点をハブとして取り除くことで得られるグラフ (hub eliminated; HE) と推論アルゴリズムのラベル伝播法 (LP) による推論の組み合せを比較した結果を**表 2** に示す．表における most freq. は最頻出語義を正解としたときの平均分類精度を現わしている．kNN，bM，DB ではパラメタ k, b を $2, 3, \ldots, 28$ の間で，HE では d を $5, 10, 15, \ldots, 95$ の間で変動させ，パラメタ k を $3, 5, 7, 9$ の間で変動させラベル伝播法に適用したときの平均分類精度の最も高い値をそれぞれ調べた．また次数制約付きグラフに対しては 3.4 節におけるグラフを連結するヒューリスティックを適用した．

⋆1 http://www.senseval.org/senseval3/data.html

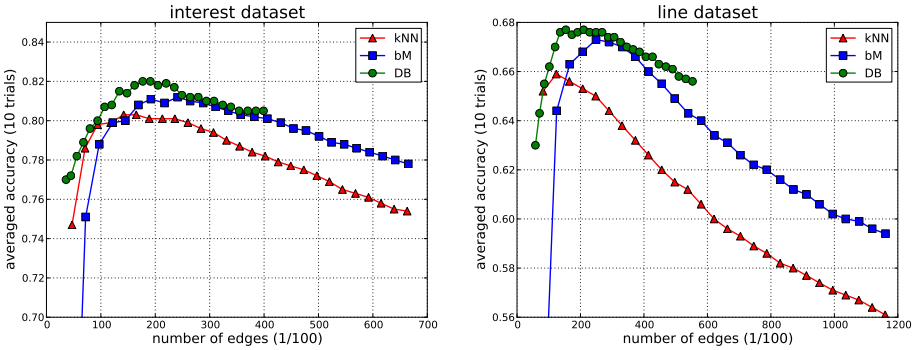

図 4 2 つの名詞 interest,line における各グラフ構築法をラベル伝播法に適用した場合の平均分類精度を比較

表 2 Pedersen の interest, line データセットにおけるデータセットのサイズと平均分類精度

	Dataset			Averaged accuracy (10 trials)				
noun	train	test	class	most freq.	LP+kNN	LP+bM	LP+DB	LP+HE
interest	236	2132	6	52.8	80.3	81.2	**82.0**	81.0
line	414	3732	6	53.5	65.9	67.3	**67.7**	67.1

パラメータ k, b を変化させ，構築されたグラフの辺の数に対する分類性能の比較の一部を**図 4** に示す．各グラフ構築法を比較し，interest, line いずれのデータセットにおいても次数制約付きグラフをラベル伝播法に用いた場合の結果が最も優れた分類性能を示している．

Senseval-3 タスクに対して，個々の名詞についてグラフ構築法を比較した結果を**表 3** に，また全 20 名詞すべての合計の平均分類精度を比較したものを**表 4** に示す．結果は Senseval-3 においても次数制約付きグラフによるグラフ構築が他の手法と比較して有効であることを示している．

Vol.2010-NL-199 No.14
2010/11/19

表 3 SE3LS の 全 20 名詞における平均分類精度とデータセットサイズ

	Dataset			Averaged accuracy (10 trials)			
noun	train	test	class	most freq.	LP+kNN	LP+bM	LP+DB
argument	32	292	5	44.0	**47.6**	47.5	47.0
arm	39	351	5	79.2	83.6	83.5	**84.6**
atmosphere	23	207	5	44.3	48.1	**48.3**	48.0
audience	29	263	4	73.2	**78.9**	78.2	78.3
bank	38	351	9	63.9	71.3	72.0	**72.7**
degree	38	343	7	63.7	71.4	73.9	**74.8**
difference	33	305	5	40.8	51.2	51.7	**52.0**
difficulty	6	62	4	26.3	36.2	35.8	**36.6**
disc	29	266	4	34.4	55.9	56.0	**56.4**
image	21	189	7	39.3	59.9	**61.3**	60.9
interest	27	249	7	36.2	55.4	55.1	**58.2**
judgment	9	85	7	27.7	36.4	34.1	**38.1**
organization	16	152	6	70.3	82.3	**84.1**	84.0
paper	32	290	7	22.6	41.8	**42.1**	41.8
party	32	297	5	66.3	72.5	73.1	**73.7**
performance	25	228	5	21.6	37.6	37.7	**38.2**
plan	24	219	3	76.0	78.7	**79.2**	79.0
shelter	29	264	4	41.5	53.4	**54.7**	53.8
sort	28	255	4	57.9	64.9	65.5	**65.6**
source	9	84	7	50.1	**53.7**	53.0	51.8

表 4 SE3LS の全名詞について，各グラフをラベル伝播法に適用した場合の平均分類精度を比較

	Averaged accuracy (10 trials)				
Initial labels	most freq.	SVM	LP+kNN	LP+bM	LP+DB
10%	51.3	59.6	61.3	61.8	**62.2**

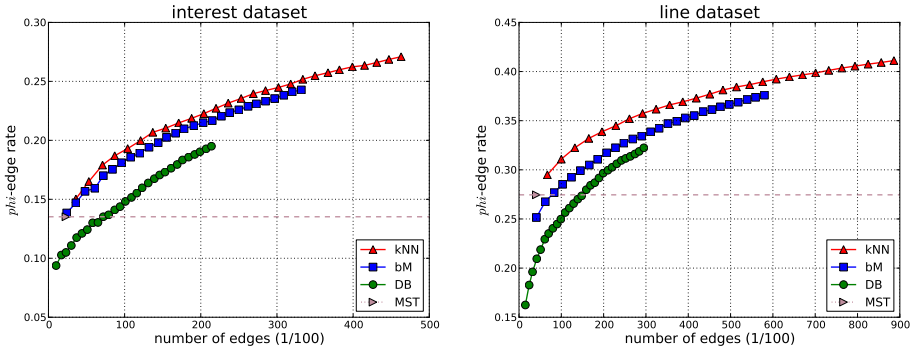

図 5 2 つの名詞 interest,line に対して，各グラフ構築法で構築されたグラフにおける ϕ-edge の割合を比較

5.3 構築されたグラフの評価

本稿の実験に用いたラベル伝播法はクラスタ仮説に基づいて設計されている．クラスタ仮説とは，似たサンプルは同じクラスラベルに割り当てられやすいという仮説である．したがって，ラベル伝播法においてはグラフがクラスタ仮説を満たしていなければ，高い分類精度を期待することができない．そこで，各グラフ構築法に対してどの方法で構築されたグラフがこの仮説をよく満たしているかを量ることで，グラフ構築法を比較する．

グラフに対するクラスタ仮説を満たす程度を量る指標を ϕ-edge を用いて定義する．**図 1** では各頂点 $v \in V$ について，接続する辺のうち ϕ-edge の割合を調べたが，本節ではグラフ G について，グラフに含まれる辺のうち ϕ-edge の割合を調べる．$l(i)$ を頂点 i に対するクラスラベルとする．以下のようにクラスタ仮説を満たす程度を量る指標を定義する．

$$\Phi(G) = \frac{\sum_{(i,j)\in E} P_{ij} 1[l(i) \neq l(j)]}{\sum_{(i,j)\in E} P_{ij}} \tag{3}$$

ただし，$1[l(i) \neq l(j)]$ は $l(i) \neq l(j)$ のときのみ 1 をとり，$l(i) = l(j)$ のときは 0 をとる指示関数である．この量はグラフ中のすべての辺において，異なるクラスラベルをむすぶ辺の割合に相当する．0 から 1 の範囲で値をとり，この量が小さければよりクラスタ仮説を満たしていると言える．この指標が最大値/最小値をとる場合の例を図 1 に示した．

いま，interest，line データセットにおいてクラスラベルがすべて既知であるとして，各手法により構築されたグラフに対して，ϕ-edge の割合を比較した結果が**図 5** である．図に

おける横軸は構築されるグラフの辺の数を示している．MST は最大全域木を構築したときの結果であり，各点がパラメタを変て構築したグラフでの結果に対応する．この結果から次数制約付きグラフを用いた場合，k-近傍グラフや b-マッチンググラフと比較して，ϕ-edge の割合はより小さくなり，クラスタ仮説を比較的よく満たしたグラフができることがわかる．

6. ま と め

半教師あり学習に用いるグラフについて，ハブに着目したグラフ構築法を提案することで，効率的な計算が可能であり，なおかつ語義曖昧性解消タスクにおいて既存のグラフ構築法と比較して，より優れた分類性能を達成することに確認した．他の手法においても次数制約付きグラフが有効であるかを確認し，どの程度ラベルありデータが不足している場合に教師あり学習と比較して有効であるかの調査を今後の課題とする．

語義曖昧性解消タスクにおけるものと同様，文書分類タスクでは文書に出現する単語を特徴量として扱うためサンプルを表現するベクトルの次元が非常に大きくなる．自然言語処理における多くのタスクでは，データの次元が高次元である場合が多く，本研究における提案手法が他の自然言語処理のタスクにおいても有効であると考えられる．本稿で提案するグラフ構築法を用いた半教師あり学習のアプローチが他の自然言語処理のタスクにおいても有効であるかを今後調査したい．

参 考 文 献

1) Dagan, I. and Itai, A.: Word sense disambiguation using a second language monolingual corpus, *Computational Linguistics*, Vol.20, No.4, pp.563–596 (1994).
2) Vickrey, D., Biewald, L., Teyssier, M. and Koller, D.: Word sense disambiguation for machine translation, *Proc. of HLT-EMNLP-2005*, pp.771–778 (2005).
3) Leacock, C., Miller, G.A. and Chodorow, M.: Using corpus statistics and WordNet relations for sense identification, *Computational Linguistics*, Vol.24, No.1, pp. 147–165 (1998).
4) Zhu, X., Ghahramani, Z. and Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions, *Proc. of ICML-2003*, pp.912–919 (2003).
5) Subramanya, A. and Bilmes, J.: Entropic graph regularization in non-parametric semi-supervised classification, *Proc. of NIPS-2009*, pp.1803–1811 (2009).
6) Wang, J., Jebara, T. and Chang, S.: Graph transduction via alternating minimization, *Proc. of ICML-2008*, pp.1144–1151 (2008).
7) Niu, Z.-Y., Ji, D.-H. and Tan, C.L.: Word sense disambiguation using label propagation based semi-supervised learning, *Proc. of ACL-2005*, pp.395–402 (2005).
8) Pham, T.P., Ng, H.T. and Lee, W.S.: Word sense disambiguation with semi-supervised learning, *Proc. of AAAI-2005*, pp.1093–1098 (2005).
9) Goldberg, A.B. and Zhu, X.: Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization, *Proc. of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pp. 45–52 (2006).
10) Celikyilmaz, A. and Hakkani-Tur, D.: A graph-based semi-supervised learning for question semantic labeling, *Proc. of NAACL-HLT-2010 Workshop on Semantic Search*, pp.27–35 (2010).
11) Maier, M., Luxburg, U.V. and Hein, M.: Influence of graph construction on graph-based clustering measures, *Proc. of NIPS-2009*, pp.1025–1032 (2009).
12) Jebara, T. and Shchogolev, V.: B-Matching for spectral clustering, *Proc. of ECML*, pp.679–686 (2006).
13) 小嵜耕平，小町　守，新保　仁，松本裕治：半教師あり語義曖昧性解消のためのグラフスパース化，情報処理学会第 196 回自然言語処理研究会，Vol.2010-NL196, No.19 (2010).
14) Alexandrescu, A. and Kirchhoff, K.: Data-driven graph construction for semi-supervised graph-based learning in nlp, *Proc. of HLT-NAACL-2007* (2007).
15) Jebara, T., Wang, J. and Chang, S.-F.: Graph construction and b-matching for semi-supervised learning, *Proc. of ICML-2009*, pp.441–448 (2009).
16) Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods, *Proc. of ACL-1995*, pp.189–196 (1995).
17) Li, H. and Li, C.: Word translation disambiguation using bilingual bootstrapping, *Computational Linguistics*, Vol.30, No.1, pp.1–22 (2004).
18) Mihalcea, R., Chklovski, T. and Kilgarriff, A.: The Senseval-3 English lexical sample task, *Proc. of Senseval-3: the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp.25–28 (2004).
19) Tabei, Y., Uno, T., Sugiyama, M. and Tsuda, K.: Single versus multiple sorting in all pairs similarity search, *Proc. of ACML-2010* (2010). (to appear).
20) Chen, J., Fang, H.-r. and Saad, Y.: Fast approximate kNN graph construction for high dimensional data via recursive lanczos bisection, *Journal of Machine Learning Research*, Vol.10, pp.1989–2012 (2009).
21) Beygelzimer, A., Kakade, S. and Langford, J.: Cover trees for nearest neighbor, *Proc. of ICML-2006*, pp.97–104 (2006).
22) Ram, P., Lee, D., March, W. and Gray, A.: Linear-time algorithms for pairwise statistical problems, *Proc. of NIPS-2010* (2010). (to appear).
23) Szummer, M. and Jaakkola, T.: Partially labeled classification with markov random walks, *Proc. of NIPS-2002*, pp.945–952 (2002).

24) Cesa-Bianchi, N., Gentile, C., Vitale, F. and Zappella, G.: Random spanning trees and the prediction of weighted graphs, *Proc. of ICML-2010*, pp.175–182 (2010).
25) Pedersen, T.: A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation, *Proc. of NAACL-2000*, pp.63–69 (2000).
26) Huang, B.: Loopy belief propagation for bipartite maximum weight b-matching, *Proc. of AISTATS*, pp.195–202 (2007).
27) Radovanović, M., Nanopoulos, A. and Ivanović, M.: Hub in space: popular nearest neighbors in high-dimensional data, *Journal of Machine Learning Research*, Vol.11, pp.2487–2531 (2010).
28) Lee, Y.K. and Ng, H.T.: An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation, *Proc. of EMNLP-2002*, pp.41–48 (2002).

付　　録

A.1　ラベル伝播法によるラベル推定

本節では，グラフを用いた推論アルゴリズムの中で，重み付きグラフ G について定義されるコスト関数を最適化することによりクラスラベルを推定するラベル伝播法について述べる．なお，スパース化されたグラフを $G' = (V, E', W')$ とし，4.1 節において求めた P を用いて，W' は任意の i, j に対して $W'_{ij} = W_{ij}P_{ij}$ であるする．E' は任意の i, j に対して P'_{ij} が 1 ならば,$(i, j) \in E'$ であり，0 ならば $(i, j) \notin E'$ である集合であるとする．

なお D は対角行列であり，各成分は $D_{ii} = \sum_i W'_{ij}$ により定義されるものとする．またグラフ G において，ラベルありデータに関連付けられている頂点を V_l とし，ラベルなしデータに関連付けられている頂点を V_u とする．

ラベル伝播法は，以下の式で定義されるコスト関数を最小化することにより，最適な識別関数 F を求める手法である．[4)]

$$\min_{F \in R^{|V| \times S}} \mathrm{tr}\{F^T \Delta F\} \tag{4}$$

ただし，Δ はグラフラプラシアン $\Delta = D - W'$ である．F は $|V| \times |S|$ の実数値行列であり，各成分は頂点 $v_i \in V$ とクラスラベル $s_j \in S$ との間隔を表す量を持つ．また，F をラベルあり頂点 V_l とラベルなし頂点 V_u に対応する行により分割した行列をそれぞれ F_l, F_u とする．いまラベルありデータの頂点とそのクラスラベルに対応する成分を 1 とし，それ以外を 0 として定義される $|V_l| \times |S|$ 行列を Y_l とする．上式における最適化はラベルなし頂点に $\Delta F_u = 0$ の，ラベルあり頂点に $F_l = Y_l$ の 2 つの制約の元で行なう．

この最適化により得られた F_u において，各頂点 $v_i \in V_u$ と間隔を最大とするクラスラベル $s_j \in S$ がラベルなし頂点のクラスラベルとして推定される．