様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成 25 年 5 月 28 日現在

機関番号：14603
研究種目：若手研究（B）
研究期間：2011～2012
課題番号：23700177
研究課題名（和文）　自然言語処理における全体最適化のための大規模・超並列処理
研究課題名（英文）　Large-scale and massively parallel computing for global optimization in natural language processing

研究代表者
小町　守（KOMACHI MAMORU）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：60581329

研究成果の概要（和文）：近年自然言語処理分野で注目されている手法の一つに、半教師あり学習手法がある。半教師あり学習は、少数のラベルつきデータ（シード）に加え、大規模なラベルなしデータの情報を用いることで、高精度な学習を行う手法である。本研究は、グラフ全体の構造が半教師あり手法において有効であることをいくつかのタスクで示した。

研究成果の概要（英文）：In recent years, semi-supervised learning has been receiving more and more attentions in natural language processing. Semi-supervised learning uses large-scale unlabeled data in addition to small number of labeled data (seed) to perform highly accurate statistical learning. This research project showed that the structure of graph affects the performance of semi-supervised methods in several tasks.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	2,700,000	810,000	3,510,000

研究分野：自然言語処理
科研費の分科・細目：人間情報学・知能情報学
キーワード：自然言語処理，ビッグデータ，半教師あり学習

１．研究開始当初の背景

最近自然言語処理分野で注目されている手法の一つに、半教師あり学習手法がある。半教師あり学習は、少数のラベルつきデータ（シード）に加え、大規模なラベルなしデータの情報を用いることで、高精度な学習を行う手法である。たとえば図 1 は「バラ」という事例から花の名前を抽出するパターンをコーパスから検索し、他の花の名前を獲得するブートストラップ法というアルゴリズムを示す。ラベルをつける事例を少数にすることで人手によるデータ作成のコストを激減させることが可能なだけではなく、無尽蔵

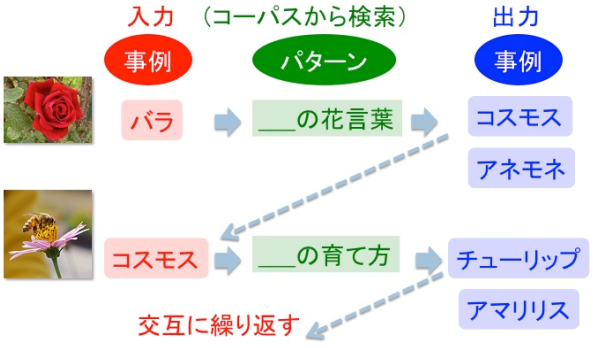

図 1: 言語処理におけるブートストラップ

に手に入るラベルなしデータの情報を組み合わせることによって、ラベルなしデータの特徴も考慮した学習を行うことができる。

申請者らはこれまでにブートストラッピング法という半教師あり手法について研究を行い、Espresso という最新のブートストラッピング法にグラフ理論的解釈を与え、自然言語処理分野タスクに対し、リンク解析の手法が応用できることを明らかにした。[小町 2010] ブートストラッピング法には反復を繰り返すにつれ初期のシード事例とは無関係な事例を獲得してしまう意味ドリフトという問題が知られているが、この意味ドリフトが Kleinberg の HITS [Kleinberg 1998] におけるトピックドリフトと同じ問題であることを示し、意味ドリフトを防ぐ手法を提案した。

それではなぜブートストラッピング法では意味ドリフトが起きてしまうのだろうか？申請者らはこの問題に対し、ブートストラッピング法が現在判明している事例と共起する局所的なパターンのみから抽出する事例・パターンを判断しているためだと予想した。また、多数の事例と類似度が高い事例（ハブ）が分類に悪影響を与えることを示し、コーパス全体から事前に計算することが可能なハブ事例の影響を抑えることで分類精度が向上することが明らかになりつつある。[小嵜 2010]

このようなグラフを用いた自然言語処理の半教師あり学習アルゴリズムとしてラベル伝播法が知られているが、申請者らは Google の大規模並列分散処理フレームワーク MapReduce を適用することにより、効率的な知識獲得手法を開発した。[小町 2010] これらの手法は 1 クラスの学習で経験的にうまく行くことが多いと分かっているが、獲得が難しいクラスが存在するという問題と、そしてどのようなシードが精度の高い意味カテゴリ学習につながるかという問題が未解決であった。

２．研究の目的

申請者らはこれまでウェブ検索クリックスルーログに対しグラフに基づくラベル伝播法を適用することにより、高精度・高再現率な意味カテゴリ知識獲得が可能であることを示した[小町 2010]が、複数カテゴリに対してグラフに基づく手法を適用する場合の理論的基盤はいまだ解明されていない。

そこで、大規模並列計算環境を用い、コーパス全体で最適化する手法を用いて複数カテゴリに対する情報抽出について理論的に検討する。具体的には、複数カテゴリのシードを用いることにより効率的にハブを発見し、コーパス全体で事例・パターンの最適化を行うことで、獲得の難しいクラスの問題を解決できるであろうという着想が背景にある。

３．研究の方法

複数クラスの分類が問題となる自然言語処理の一タスクとして語義曖昧性解消を用い、コーパス全体を見て競合するクラスのパターン・事例を負例として用いることで高精度に語義曖昧性を解消する手法の研究と理論的背景の考察を行った。

(1) 競合するクラスの事例を人手で抽出し、「ストップリスト」と呼んで獲得対象から除外する、という手法[McIntosh 2010]に対し、グラフに基づくノードのランキング手法を用いることでストップリストの自動構築を行った。また、同じグラフに基づくノードのランキング手法を半教師あり学習のシード事例選択に適用し、ラベル付け対象の教師なし事例選択を行なった。語義曖昧性解消タスクでこれらの手法の有効性を検証した。

(2) グラフに基づく半教師あり学習の課題の一つであるグラフ構築方法について、効率的で高精度な手法を提案した。具体的には、k近傍グラフという、各ノードの近くのk個のノードに対してエッジを貼るというグラフ構築法がよく知られているが、この方法ではたくさんのノードからエッジが張られるハブノードが競合するクラスのパターン・事例を仲介してしまうと考え、相互のk近傍に入っているときだけエッジを貼るという相互k近傍グラフというグラフ構築法を自然言語処理データに適用した。語義曖昧性解消タスクと文書分類タスクで評価した。

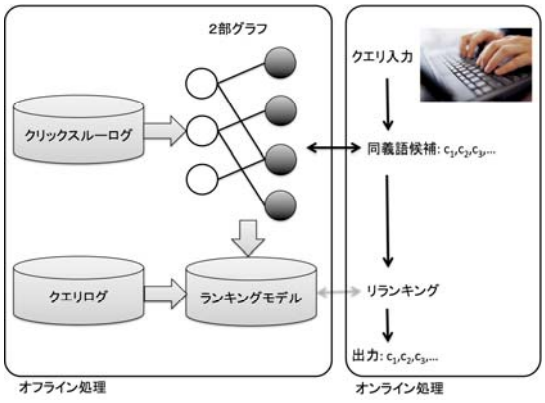

図 2: 検索クエリ・クリックスルーログからの同義語獲得

(3) グラフに基づく手法が自然言語処理以外のデータでも効果があるかどうか調べるため、図 2 に示すようにウェブ検索のクエリログとクリックスルーログからグラフを構築して同義語の獲得実験を行なった。ここではハブの影響を受けない手法として知られているグラフラプラシアンを用いた教師なし手法をベースラインとし、教師あり手法と比較した。

４．研究成果

(1) 提案するグラフのノードのランキング手法を語義曖昧性解消タスクにおけるシード選択とストップリスト構築で評価した。

評価は語義曖昧性解消で用いられる Senseval-3 コーパス中の最頻出の 7 つの名詞と line, interest データセットを用いた。評価には Mean Average Precision (MAP) と Area Under Curve (AUC) および 30 位適合率 (P@30) を用いた。

ベースラインにはランダムにシードを選んだ手法を用い、提案手法はインスタンスのランキングに HITS [Kleinberg 99] アルゴリズムを用いる。

表 1 にシード選択の比較実験結果を示す。いずれの尺度においても提案手法がランダムベースラインを上回り、グラフを用いた手法の効果が示された。

表 1: シード選択の比較実験結果

	MAP	AUC	P@30
ランダム	68.4	55.1	78.9
HITS	74.2	64.7	85.2

また、表 2 にストップリストの比較実験結果を示す。シード選択と同様、HITS アルゴリズムを用いてインスタンスのランキングを行なう手法の有効性が示された。

表 2: ストップリスト構築の比較実験結果

	MAP	AUC	P@30
ランダム	23.2	57.1	36.3
HITS	24.3	59.4	39.4

(2) 提案する相互 k 近傍グラフを語義曖昧性解消タスクと文書分類タスクにおいて評価した。

語義曖昧性解消タスクには interest データセット、line データセットを用い、文書分類タスクにおいては Reuters データセットおよび 20 newsgroups データセットを用いた。データセットの全事例中ランダムに選択した 10%をラベルあり事例、残りの 90%をラベルなし事例として半教師あり学習を行なった。データによってはグラフが連結にならないという問題を解消するために、最大全域木を求め、それぞれのグラフ構築法に組み合わせることで非連結成分が生まれないようにした。分類アルゴリズムにはグラフベースの半教師あり学習でよく用いられている local and global consistency を使用した。

ベースラインとしては広く用いられている k 近傍グラフ、入次数と出次数が同じになるという制約をかけた最新手法である b マッチンググラフ [Jebara 2009] の 2 つを比較した。評価には分類精度を用いた。

実験結果を表 3 に示す。表から分かるように、提案する相互 k 近傍グラフが最も性能が高い。また、b マッチンググラフは 20 newsgroups データセットにおいては計算時間がかかりすぎ、実験が終わらなかった。K 近傍グラフ、相互 k 近傍グラフともに計算量は $O(n^2+kn\log n)$ である（ただし n はノード数、k は近傍の数）のに対し、b マッチンググラフの時間計算量は $O(bn^3)$ であり（ただし b は次数、n はノード数）、巨大なデータセットに対しては計算量が膨大になるということが確認された。

表 3: 語義曖昧性解消タスクと文書分類タスクにおける分類精度の比較

	k 近傍グラフ	b マッチンググラフ	相互 k 近傍グラフ
interest	81.44	82.17	82.38
line	69.06	69.35	70.39
Reuters	82.60	84.42	84.85
20 newsgroups	75.19	--	75.41

(3) 構築した大規模なグラフを用いた同義語獲得手法と、それを素性に用いた識別学習による同義語獲得手法を比較した。

実験には Yahoo! Japan ウェブ検索に入力された 1 ヶ月分の検索クエリ、クリックスルーログを用いた。

ベースラインとしてはグラフラプラシアンを用いてラベル伝播を行なう [小町 2011] を用い、比較手法として ListNet という識別モデルによるランキング学習を行なった。ListNet の素性として、ベースラインと同様の素性を用いたもの、素性テンプレートを用いたもの、そして教師なしで素性抽出を行なう非線形拡張を行なったもの、の 3 種類を

比較した。

同義語獲得タスクにおける各手法の 1 位正解率を表 4 に示す。結果から分かるように、識別学習を行なった手法は先行研究の手法より高い精度を示した。また、識別学習することによって柔軟な素性を用いることができ、さらに正解率を向上させることができた。最後に、非線形拡張を行なうことで、素性エンジニアリングの問題を回避しつつ高い正解率を達成することができた。

表 4: 同義語獲得タスクにおける 1 位正解率

	1 位 正解率
ノイジーチャネルモデル	0.557
ListNet （ノイジーチャネルモデル素性）	0.584
ListNet （素性テンプレート）	0.655
ListNet （非線形拡張）	0.735

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① 小嵜耕平, 新保仁, 小町守, 松本裕治. 相互 k-近傍グラフを用いた半教師あり分類. 人工知能学会論文誌, 28 巻 4 号, 2013. 査読有（掲載確定）
② 内海慶, 小町守. ウェブ検索クエリログとクリックスルーログを用いた同義語獲得. 情報処理学会論文誌: データベース（TOD56）, 6 巻 1 号, 16-28, 2013, 査読有

〔学会発表〕（計 3 件）

① Kei Uchiumi, Mamoru Komachi, Keigo Machinaga, Toshiyuki Maezawa, Toshinori Satou and Yoshinori Kobayashi. Japanese Abbreviation Expansion with Query and Clickthrough Logs. In Proceedings of the 5th International Joint Conference on Natural Language Processing. 2011.11.10. Thailand.
② Kohei Ozaki, Masashi Shimbo, Mamoru Komachi and Yuji Matsumoto. Using the Mutual k-Nearest Neighbor Graphs for Semi-supervised Classification on Natural Language Data. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning. 2011.6.23. USA.
③ Tetsuo Kiso, Masashi Shimbo, Mamoru Komachi and Yuji Matsumoto. HITS-based Seed Selection and Stop List Construction for Bootstrapping. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011.6.21. USA.

〔産業財産権〕
○取得状況（計 1 件）

名称：適正単語取得装置、機械学習装置及び方法
発明者：小町守, 颯々野学
権利者：同上
種類：特許
番号：特許第 5042268 号
取得年月日：24 年 7 月 20 日
国内外の別：国内

〔その他〕
ホームページ等
http://cl.sd.tmu.ac.jp/

６．研究組織
(1)研究代表者

小町　守（KOMACHI MAMORU）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：60581329