様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　6 月　4 日現在

機関番号：14603
研究種目：基盤研究(C)(一般)
研究期間：2009～2011
課題番号：21500141
研究課題名（和文）
リンク解析に基づく自然言語処理・テキストマイニング技術の開発
研究課題名（英文）
Link Analysis Approaches to Natural Language Processing and Text Mining
研究代表者
新保 仁（SHIMBO MASASHI）
奈良先端科学技術大学院大学・情報科学研究科・准教授
研究者番号：90311589

研究成果の概要（和文）：リンク解析技術を活用した，自然言語処理テキストマイニング技術を開発した．ある種のリンク解析法が用いる，2 節点間の全経路を考慮した類似度計算法にヒントを得て，自然言語文を構文解析する際に障害となっている，文内の並列構造の解析（並列構造の有無・範囲同定）技術を考案した．さらに，自然言語データをネットワーク（グラフ）として表現し，リンク解析手法を適用する際に障害となる「ハブ」（数多くのグラフ節点と結合した節点）の影響を調査し，その悪影響を軽減するグラフ構築法を提案した．

研究成果の概要（英文）：We investigated how to effectively apply link analysis techniques to the tasks of natural language processing. Specifically, we developed a method for analyzing conjunctive structures in natural language sentences. The method borrows an idea from the similarity computation method in link analysis, which counts all paths between the nodes of interest. Our proposed method applied the idea to finding the most similar phrases in a sentence, which are likely to form coordination. We also proposed a graph construction method that reduces the influence of "hubs", which are the nodes located close by to many other nodes in a graph.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,300,000	390,000	1,690,000
2010 年度	1,100,000	330,000	1,430,000
2011 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：リンク解析
科研費の分科・細目：知能情報学
キーワード：自然言語処理，データマイニング，テキストマイニング

１．研究開始当初の背景

リンク解析（あるいはネットワーク解析）は，Web の検索技術として注目を集め，その後，多くの研究分野に影響を与えている．しかしながら，自然言語処理・テキストマイニング分野においては，重要節点計算法の PageRank といった既存の手法をほぼそのまま文書要約（重要文抽出）などの限られたアプリケーションに適用する，といった応用にとどまっていた．リンク解析には，節点重要度計算以外にも有効と考えられる手法が数多く存在するが，それら技術が十分活用されているとは言いがたい状況であった．

２．研究の目的

リンク解析技術を活用した，新しい自然言語処理・テキストマイニング手法を開発する．テキストデータをグラフ（ネットワーク）として表現し，リンク解析手法を用いることで，効率的かつ効果的な処理を行う．さらに，リンク解析と機械学習技術の組み合わせによって，これまで実用化されてこなかった新しい自然言語処理アプリケーションを開拓する．

３．研究の方法

実用的な研究意義を維持することは重要であり，様々な自然言語処理タスクを対象に評価を行うことで，リンク解析技術を適用することが妥当なタスクの開拓を目指した．そういったタスクを対象に，問題点があれば，より効果的な適用方法を考案する，という方法を取った．

内容については研究成果欄にて述べるが，研究期間の前半は自然言語文の並列構造解析，後半はグラフ構築やグラフに基づく半教師あり学習法（特にセルフトレーニング法）における「ハブ」（多くの事例の近傍に位置する事例―高次元データでは多くの場合に出現する）の影響に着目して研究を行った．

４．研究成果

(1) 文内の並列構造解析法の開発
自然言語文の構文解析精度の向上は目覚しいものがある．しかしながら，文中の並列構造（「A と B と C」などの構文構造）は，並列をなす単位と，構文木の構造単位が必ずしも一致しないこともあって，いまだに構文解析の障害として残っている．
研究代表者は，平成 20 年度まで，グラフ中の全経路を枚挙して類似度として用いるリンク解析手法を応用し，文中の文字列間の全アラインメントに基づいて，文中の並列範囲を推定する技術を開発してきた．平成 21 年度は，この並列句解析技術を発展させ，これまで英語のみを対象に開発してきた解析手法を，日本語並列句および，ネストした並列句の解析，という，より困難な問題に適用した．これによって，提案手法の応用可能範囲を広げ，さらに，これらの問題におけるアラインメントグラフ上のコスト学習法の実用性を示したことが主な成果であ

る.

具体的には，アラインメントグラフに機械学習法の一種である構造化パーセプトロン(structured perceptron)を組み合わせ，日本語文においても並列句同士の類似度が適切に学習できることを実証した．英文の並列句がほぼ間違いなく「and」「or」といった並列句マーカーを伴うのに対し,日本語の並列句は「と」といったマーカーは必ずしも並列句を導くとは限らない.例として,「清水寺と二条城に行った」,「友達と二条城に行った」はいずれも助詞「と」を含む．しかしながら，前者は並列句を含むのに対し，後者は含まない.このように日本語並列句解析は,「並列句の範囲同定」に加え「並列句が存在するか否かの判定」という処理も必要とされ,問題はより困難である.我々はこのような問題に対しても，グラフ形状を工夫する（バイパスと呼ばれる経路を追加する)ことで対処した.

また，並列句はしばしばネストして文中に出現するが，従来のアラインメントグラフを用いる並列句解析法ではこのような文を扱うことが不可能であった．我々は，このような文についても，アラインメントグラフとともに構文規則に基づく制約を併用し，効率良く学習・解析が可能であることを示した.

(2) リンク解析法の関連性に基づくセルフトレーニング技術の改良

自然言語処理においては，必要な言語資源(事例）を，人手の介在を極力排して効率良く獲得するために，セルフトレーニング(self training; あるいはブートストラッピング bootstrapping)と呼ばれる手法がしばしば用いられる．我々は Espresso ブートストラッピング法と，Kleinberg による HITS と呼ばれるグラフ節点の重要度算出（ランキング）手法との類似性を指摘した.Espresso も含めたブートストラッピング・セルフトレーニング法では，アルゴリズムを繰り返すにつれ，本来獲得したいクラス以外のオブジェクトばかりが獲得されてしまう問題（意味ドリフト）が知られているが,我々の研究は,ドリフトがでたらめな方向に向かって起きるのではなく，事例獲得の過程をグラフとして見なした場合，HITS ランキングが上位事例ばかりを獲得するバイアスが存在すること（リンク解析分野では「トピックドリフト」と呼ばれる類似の現象が知られている)を示唆している.

我々はさらに，この観察結果を，意味ドリフトの予防に積極的に活用する方法を提案し，その有効性を実験によって示した.すなわち，HITS ランキングが高い事例を予め「ストップリスト」として排除することで，意味ドリフトはある程度予防できる.なお，HITS 法で上位にランキングされる事例は，後述の「ハブ」の一種とみなすことができ，従来その負の側面ばかりが指摘されてきたが，本研究では，これを機会学習法で言うところの「負例」の一種とみなして活用したことになる.

(3) グラフ構築への応用

自然言語処理における大きな問題の一つに,ラベル付きデータ（正解データ）を作成するためのコストが挙げられる．少数のラベル付きデータを仮定する，半教師あり学習法はその克服のための有望な手法であり,なかでも，グラフに基づく半教師あり法(Zhu et al. および Zhou et al. の各種ラベル伝搬法など）の有効性に注目が集まっている．これらのグラフに基づく手法は,

データがグラフとして表現されていることを仮定しているため，自然言語データなどの非グラフデータに適用する際には，一旦データをグラフに変換しなければならない．そのための手法としてはもっぱら，k-近傍グラフ（データ点各々について,最も類似するk 個のデータ点を辺で結ぶ)が用いられる．我々はこのグラフ変換法を自然言語データ(語義曖昧性解消用のベンチマークデータ）に適用し，評価・分析を行った．その結果,少数の「ハブ」と呼ばれる数多くの節点と接続された節点が生成され,後続のリンク解析に基づく半教師あり学習の精度に悪影響を与えることがわかった.その原因は k-近傍関係が非対称であることにあり，一方，相互ｋ近傍グラフと呼ばれるグラフが相対的にハブを軽減することを発見した．このグラフ構築法には，グラフの連結性が失われやすいという欠点があるが，これは最大全域木などと組み合わせることで容易に克服することができる.

自然言語データをはじめとする高次元データでは，本質的にハブが存在する可能性が高いこと（「次元の呪い」と総称される現象の一つ）が報告されている(Radovanovic et al. 2010)． このため，語義曖昧性解消以外の多くのタスクでハブによる悪影響が見られると予想され，このグラフ構築法はこれらにも有効であると考えられる

以上の(1)-(3)における成果は，いずれも，ACL，CoNLL といった自然言語処理関連の主要な国際会議に採録され，その有用性を認められている.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計６件）

(1) Amin Mantrach, Nicholas van Zeebroeck, Pascal Francq, Masashi Shimbo, Hughes Bersini, and Marco Saerens.
Semi-supervised classification and betweenness computation on large, sparse, directed graphs.
Pattern Recognition, Vol. 44 (2011), pp.1212-1224. 査読あり.

(2) Silvia Garcia-Diez, Francois Fouss, Masashi Shimbo, and Marco Saerens.
A sum-over-paths extension of edit distances accounting for all sequence alignments.
Pattern Recognition, Vol. 44 (2011), pp. 1172-1182. 査読あり.

(3) Amin Mantrach, Luh Yen, Jerome Callut, Kevin Francoisse, Masashi Shimbo, and Marco Saerens.
The sum-over-paths covariance kernel: a novel covariance measure between nodes of a directed graph.
IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32 (2010), pp.1112-1126. 査読あり.

(4) 原 一夫, 新保 仁, 松本 裕治.
文法制約と系列アラインメントによる並列構造の解析.
人工知能学会論文誌, Vol. 25 (2010), pp.560-569. 査読あり.

(5) 大熊秀治, 原一夫, 新保仁, 松本裕治.
バイパス付きアラインメントグラフを用いた日本語並列句検出と範囲同定.
人工知能学会論文誌 Vol. 25 (2010), pp. 206-214. 査読あり.

(6) 小町守, 工藤拓, 新保仁, 松本裕治.
Espresso 型ブートストラッピング法における意味ドリフトのグラフ理論的分析.
人工知能学会論文誌 Vol. 25 (2010), pp. 233-242. 査読あり.

〔学会発表〕(計 6 件)

(1) Kohei Ozaki, Masashi Shimbo, Mamoru Komachi, and Yuji Matsumoto.
Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language data.
In: *Proceedings of the 15th Conference on Natural Language Learning (CoNLL 2011)*.
2011 年 6 月 23 日. Portland, OR, USA. 査読あり.

(2) 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 (ACL 2011): Short Papers*.
2011 年 6 月 21 日. Portland, OR, USA. 査読あり.

(3) Silvia García Díez, François Fouss, Masashi Shimbo, Marco Saerens. Normalized sum-over-paths edit distances.
In: *Proceedings of the 2010 International Conference on Pattern Recognition (ICPR 2010)*.
2010 年 8 月 24 日. イスタンブール. 査読あり.

(4) Kazuo Hara, Masashi Shimbo, Hideharu Okuma, and Yuji Matsumoto.
Coordinate Structure Analysis with Global Structural Constraints and Alignment-Based Local Features.
In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009)*.
2009 年 8 月 5 日.シンガポール. 査読あり.

(5) Hideharu Okuma, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto.
Bypassed Alignment Graph for Learning Coordination in Japanese Sentences.
In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009): Short Papers*.
2009 年 8 月 4 日. シンガポール. 査読あり.

(6) Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto.
A Graph-based Approach for Biomedical Thesaurus Expansion.
In: *Proceedings of the Third ACM International Workshop on Data and Text Mining in Bioinformatics (DTMBIO)*.
2009 年 11 月 6 日. 香港. 査読あり.

６．研究組織

(1)研究代表者

新保 仁（SHIMBO MASASHI）
奈良先端科学技術大学院大学
・情報科学研究科・准教授
研究者番号：90311589

(2)研究分担者

なし

(3)連携研究者

なし