

論文内容の要旨

博士論文題目 Graph-Theoretic Approaches to Minimally-Supervised Natural Language Learning
(グラフ理論的観点からの自然言語処理における弱教師あり学習)

氏名 小町 守

(論文内容の要旨)

自然言語処理におけるブートストラッピング手法は、人手による資源作成のコストを減らすための弱教師あり学習の一つである。ブートストラッピング手法は、獲得対象となるクラス、たとえば is-a 関係の少量のインスタンス (例 (cat, animal)) をシードとして与え、コーパスからインスタンスと共起するパターン (例 “X such as Y”) を抽出する。抽出された共起パターンは、そのパターンと共起する別のインスタンスを獲得するために用いられる。このようなステップを反復的に適用することで、人手による介入を最小限にシードインスタンスの数を簡単に増やすことができる。ブートストラッピング手法の考え方は関係抽出や固有表現認識などさまざまな自然言語処理のタスクに応用されている。しかしながら、ブートストラッピング手法には、反復が進むにつれシードインスタンスとは無関係なインスタンスを抽出してしまう、意味ドリフトと呼ばれる傾向がある。

本研究の主な貢献は、Espresso 型のブートストラッピング手法における意味ドリフト現象が、Kleinberg の HITS と同じ原因によって引き起こされているのだという事実を、グラフ理論を用いて定式化し、単純化されたブートストラッピング手法を用いて解析的に示したことである。また、グラフ理論に基づく2つのアルゴリズム (von Neumann カーネルと正則化ラプラシアン) が意味ドリフトの影響を軽減することを、いくつかの自然言語処理タスクによって検証した。提案手法は Espresso より高い精度を達成し、さらに、先行研究より設定すべきパラメータ数が少ないため、調整が容易である。

本論文ではまずブートストラッピング手法について概観する。Espresso という最新のブートストラッピング手法について詳述し、実験的な性能を評価する。Espresso は他のブートストラッピング手法より高い性能であることを示し、Espresso に対して3つの修正を行なうことでさらに性能が高い Tchai という Espresso 型の新しいアルゴリズムを導く。そして Espresso 型のブートストラッピングアルゴリズムに対してグラフ理論に基づく分析を行ない、トピックドリフトと意味ドリフトの関係について示し、2つのグラフに基づく手法を提案する。最後に提案手法を3つの自然言語処理タスクに適用した結果について報告する：語義曖昧性解消、対訳辞書構築、そして意味カテゴリ学習である。実験結果によって、提案手法の一つである正則化ラプラシアンは Espresso と同等の性能を持ち、調整が簡単で、大規模データにもスケラブルであることを示す。

氏名	小町守
----	-----

(論文審査結果の要旨)

平成22年1月20日に開催した公聴会の結果を参考に平成22年2月16日に本博士論文の審査を行った。以下のとおり、本博士論文は、提案者が独立した研究者として、研究活動を続けていくための十分な素養を備えていることを示すものと認める。

小町守は、本博士論文において、ブートストラッピングに基づく知識獲得手法である Espresso アルゴリズムをグラフ理論に基づく重要性・関連性の計算モデルに基づいて定式化し、その性質の理論的および実験的な性質を明らかにした。また、グラフ理論に基づく手法を利用し、ブートストラッピング法の弱点である意味ドリフト現象を軽減できる知識獲得手法を提案した。本論文の貢献は次のようにまとめることができる。

1. 知識獲得のためのブートストラップ法としてよく知られた Espresso アルゴリズムの性質をグラフ理論に基づく手法により説明した。特に、一般的な知識抽出規則が学習されてしまう意味ドラフトという現象が、リンク解析で知られる HITS などと言われるトピックドリフトと同じ原因によることを示した。
2. グラフ理論に基づくリンク解析手法で用いられるフォン・ノイマンカーネルと正則化ラプラシアンカーネルが意味ドリフト現象を緩和することをいくつかの自然言語処理タスクを用いて実験的に示した。
3. 上記提案にしたがってブートストラップに基づく知識獲得アルゴリズムを提案し、Espresso に比べて調整すべきパラメータ数が少ないにも関わらず、より高い性能を達成することを示した。

以上のように、ブートストラップ法に対する理論的考察を行い、代表的アルゴリズムである Espresso の弱点の原因を説明し、さらに、より精度の高い手法を提案した本研究は、独創性が高く、しかも実用的であり、自然言語処理の分野において高い貢献があると評価する。

よって、本論文は、博士(工学)の学位論文として価値あるものと認める。