

審査の結果の要旨

氏 名 張 耀中

自然言語処理において、語彙化文法を用いた深い統語解析は、テキストに対し詳細な統語・意味情報を付与する手法の1つであり、分野における非常に重要な技術である。その中でsupertagging は、深い統語解析の速度を向上させる上で鍵となる方法である。高精度の supertagger は統語解析の第一ステップである語彙項目の選択における曖昧性を劇的に減少させるため、深い構文解析のための前処理手法として広く使用されている。一方で、supertagging はテキストに詳細な構文情報を付与するための効率的な手法として、機械翻訳など、統語解析以外の自然言語処理タスクへの応用も可能である。本論文では、大規模な実テキストに対し高精度かつ高速な深い統語解析を行うことを目的とし、supertagging の構造的学習法について研究を行なったものである。

本論文の目標とするSupertagger の学習に関しては、2つの困難な問題が知られている。1つは、supertag は文法から抽出されることから、タグ集合が巨大になり、かつ supertag の選択は文内で遠く離れた位置にある単語やその品詞などに依存することから、正しい supertag の予測が難しくなるというものである。遠く離れた位置にある単語の情報を単純にモデルに組み込むと計算量が膨大となるため、従来の手法ではそれらの依存性は無視されることが多い。この計算量の増加の問題は、タグ集合が巨大な場合には特に深刻である。もう1つの問題は、supertagger の学習は一般に構文解析器からは独立して行われるため、supertagger の学習における目的関数が構文解析器の学習における目的関数と異なっており、このことが最終的な構文解析の結果に悪影響を及ぼす可能性があるということである。

本論文では、supertagger の学習において文の構文構造を考慮することで両問題を解決している。まずはじめに、supertagging を従来と同じく系列ラベリングの問題として扱い、長距離の依存関係は無視することがどの程度 supertagging の結果に影響するのかを精査している。この実験結果により、長距離の依存関係の情報が supertag の曖昧性解消にとって重要であることが解明された。それを基に、1つ目の問題に対する解決法として、依存文法の枠組みで長距離の依存関係をモデル化することで、そのような依存関係を効率的に supertagger の学習に統合する手法を提案している。語彙化文法のひとつである主辞駆動句構造文法 (HPSG) における supertagging では、主語・補語

に関する統語情報が重要であり、依存文法の枠組みは、この情報を表す近似表現であると捉えることができるため、依存文法による構文解析結果を素性としてモデルに取り込むことで、長距離の依存関係を考慮することができる。本手法では長距離の依存関係の情報を supertagging におけるソフトな制約としてモデルに組み込むため、計算量的複雑さを抑えながら supertagger の精度を向上させることができる。

さらに、2つ目の問題に対する解決法として、supertagger の学習における目的関数を構文解析器の学習における目的関数へと近付けることを狙い、構文解析器を supertagger の学習時に制約として用いる手法を提案している。この手法では、元の HPSG を近似する文脈自由文法 CFG を用いて最良優先探索を行うことで、学習時にモデルのパラメータを更新する際に文法の制約を満たす supertag のみを用いる。この手法により、標準的な評価セット (Penn Treebank Section 23) に対して、ベースラインの手法に比べ述語項構造認識の F 値で 0.68% の改善が得られ、また、現在最高精度を持つ HPSG 構文解析器よりも高速であり、かつ遜色のない 89.31% の精度を得る成果を得ている。さらに両方の手法を適用することについても検討が加えられている。

なお、本論文の成果は共同研究者とともに得たものであるが、申請者が主体的に取り組み得たものであることを確認している。

以上をまとめるに、本論文は情報理工学におけるコンピュータ科学、特に自然言語処理において重要な貢献をしたと判断できる。よって本論文は博士 (情報理工学) の学位請求論文として合格と認められる。