論文の内容の要旨

Structure-guided Supertagger Learning

(構造を利用したSupertagger の学習)

氏名　張　耀中

Deep syntactic analysis by lexicalized grammar parsing is an important task in the field of natural language processing (NLP), since it provides rich syntactic and semantic description of text data. Supertagging is a key speed-up technique for deep parsing. An accurate supertagger can greatly reduce the lexical ambiguity for the downstream parser and it is widely used as a pre-process module in deep parsing. On the other hand, supertagging can also be used as an effective way to provide syntactic information for other NLP tasks, such as machine translation. To make an accurate and fast deep parser for large scale real world data, this thesis focuses on the research of structural learning problems for supertagging task.

In supertagger learning, there exist two challenging problems. First, supertags are usually derived from a complex grammar. This results in a large tag set and makes the supertag prediction relying on the information (e.g., words, part-of speech tags and other supertags) which is long-distance away in a sentence. This long-distance information is commonly ignored in traditional way of supertagger training, since the incorporation of long-distance information into traditional models brings much computational cost. The problem is especially severe when the tag set is large, because of the exponential growth of the model complexity. Second, supertaggers are usually trained separately from the parser. This pipeline parsing strategy poses a problem that the training objective of a supertagger deviates from the final parser (i.e., different loss functions), which will harm the performance of the final parser.

We solve these two problems by incorporating structural guidance in supertag learning process. In detail, we first investigate supertagging from the traditional point of view (as a sequence labeling task) and examine to what extent the ignorance of long-distance information in the sequence labeling formulation affects the supertagging results. As the experiments confirmed, the long-distance information is crucial for supertag disambiguation. For the first problem, we propose an effective method by modeling this long-distance information in dependency formalism

and integrate it into the supertagger training process. For HPSG supertags, subject/complement slots carry major syntactic information inside the supertags. The dependency formalism can be treated as an approximated representation for subject/complement information. Therefore, we first model the dependency relationships between words and use them to generate long-distance features for supertag disambiguation. This approach incorporates long-distance information as soft constraints for supertagging, which increases the accuracy of the supertagger while keeping the computational complexity tractable.

To address the second problem, we propose an on-line forest-guided training method to make the training objective of a supertagger closer to that of the parsing task. We use a CFG grammar to approximate the original HPSG grammar and apply best-first search to select grammar-satisfying supertag sequences for the parameter updates in the training stage. On the standard test set (Penn Treebank Section 23), we achieved an absolute 0.68% improvement in the F-score for predicate-argument relation recognition and got a competitive result of 89.31% with a faster parsing speed, compared to a state-of-the-art HPSG parser.

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ひとつであり，自然言語処理（NLP）の分野における重要な技術である．Supertaggingは，深い統語解析の速度を向上させる上で鍵となる手法である．高精度の supertagger は統語解析の最初の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NLPタスクへの応用も可能である．
本論文では，大規模な実テク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う．

Supertagger の学習に関して，二つの困難な問題が知られている．一つは，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．
もう一つの問題は，supertagger の学習は一般に構文解析器とは独立に行われるため，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．

本研究では，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す

る．まず始めに，supertagging を従来と同じく系列ラベリングの問題として扱い，長距離の依存関係を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する．この実験の結果より，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分かった．次に、一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する．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．

二つ目の問題に対する解決法としては，supertagger の学習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，構文解析森を supertagger の学習時に制約として用いる手法を提案する．この手法では，元の HPSG を近似する 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に，なおかつ遜色のない 89.31% の精度を得ることができた．