様式：課程博士用

2011年　8月　22日

電子・情報工学専攻	学籍番号	021705
申請者 氏名	山本　悠二	

指導 教員	増山繁 石田好輝

論　文　要　旨　(博士)

論文題目	語彙的・構文的な優先規則を考慮した統計的日本語係り受け解析についての研究

(要旨　1,200字程度)

日本語係り受け解析は自然言語処理における基本技術として認識されている．現在，日本語係り受け解析は，ルールベースに基づく方法から，係り受け情報が付与されたコーパスから機械学習を用いて係り受け解析を行う統計的日本語係り受け解析へと移行しつつある．統計的日本語係り受け解析は「Shift-Reduceによる決定的な解析手法」と「相対的な比較による解析手法」に大別される．前者は，入力文節数に比例した時間で係り受け解析ができる反面，長距離の係り先の同定が苦手であるという特徴がある．また，後者は，長距離の係り先の同定に有利である反面，係り受け解析時間が入力文節数の二乗に比例するため，それほど高速に動作しない．

本論文では統計的係り受け解析について次の3点に関する研究成果を示す．

1. 入力文節数の二乗の時間計算量であるアルゴリズムを用いて，従来法の相対的な比較による解析より，高い係り受け解析性能を実現させる研究
2. 入力文節数に対して解析時間が線形の傾向を保ちつつ，従来法のShift-Reduce法に基づく決定的な解析より，高い係り受け解析性能を実現させる研究
3. 係り受け解析における半教師あり学習を想定して，① オンライン学習であり，かつ，② ラベル付きデータとして与えられた解析結果に誤りを含んでいる場合においても頑健に学習できる手法を構築するための基礎的研究

1番目の研究において，従来法では文節間距離がカテゴリ分けされていることから，複数の係り先候補が同一の文節間距離カテゴリに属するとき，距離による弁別ができないことを指摘した．これは「可能な範囲で，できるだけ近い係り先を優先する」というヒューリスティクスを反映できない場合があることを表している．提案手法では，係り先文節候補集合から二つの文節候補を取り出すときに，それぞれの文節候補に対して係り元に近い／遠いという情報を素性として追加して係りやすさを求める．実験では，京都テキストコーパス4.0を用い，ベースライン手法と比べて係り受け正解率，文正解率が有意に改善されていることが確認された．

2番目の研究において，Shift-Reduceによる決定的な解析と相対的な比較による解析とを同一の識別モデルを用いて組合せる手法を提案した．具体的には，「曖昧性が生じる係り受けは，係り元，係り先候補に特徴がある」ことを利用し，曖昧性があるような係り受けに関して決定的な係り受けでは係り先の同定を保留しておき，後に係り先の相対的な比較によって係り先を定めるというものである．京都テキストコーパスを用いて提案手法を比較したところ，係り先候補の比較に基づく解析方法の1つである相対モデ

ルと比較してほぼ同等の解析性能を持ち，かつ，実行時間が 2.4 倍程度高速であることを示した．

3 番目の研究において，まず，半教師あり学習のベースとなるクラスタリング手法について，オンライン学習による方法を提案し，性能評価を行った．次に，先のクラスタリング手法から派生して，ラベル付きデータに誤りが含まれているデータセットにおいても頑健に半教師あり学習を行う手法を検討し，隣接文節対の係り受け関係の可否判定にその手法を適用した．提案した半教師あり学習では，逐次的に与えられるデータ点について，ラベルの有無にかかわらず，そのデータ点の分離超平面からのマージン距離が十分大きい(1 以上)ときは重みベクトルを更新しないというものである．このようにすることで，分類器が正しく予測できている場合に学習事例に対して重みベクトルを無理に更新することを避けられることが期待できる．そして，隣接文節の係り受け判定の問題を対象にして提案手法の性能評価を行った．特に，ラベル付きデータにノイズを加えた場合において提案手法の有効性が確認された．