平成26年度 情報工学コース卒業研究報告要旨

石川研究室	氏　　名	鈴木　寛大
卒業研究題目	半構造データマイニングを用いた 構文木コーパスの誤り自動訂正に関する研究	

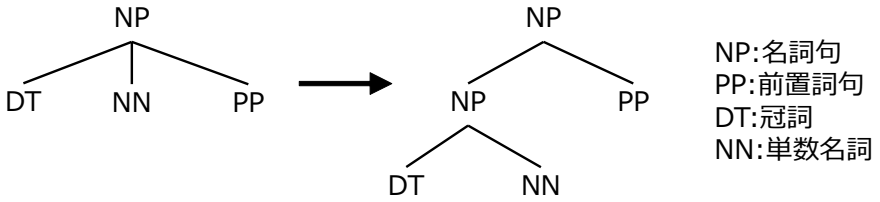
さまざまな情報をタグとして付与した自然言語文の集積体であるタグ付きコーパスは，自然言語処理技術の研究・開発の資源として利用されている．タグ付きコーパスの作成には人手が介在するため，誤りの混入は避けられず，コーパスの質を低下させる原因となる．

これに対し，コーパス内の誤りを検出・訂正する手法が提案されている．例えば，加藤らは構文木コーパスの誤り訂正手法を提案している．この手法では，コーパス中に複数回出現する単語列に，出現箇所によって異なる構文構造が割り当てられている場合，二つの構文構造の内，出現頻度の高い方が正しく，低い方が誤りを含んでいると判断し，後者を前者へ変換するルールを作成する．しかしこの手法では，コーパス中に誤った構文構造が存在しても，それが覆う単語列がコーパス中の別の箇所に出現していなければ，誤りを検出することができず，誤り訂正ルールも抽出できない．

そこで本研究では，単語列を参照することなく構文構造の誤りを検出し，誤り訂正ルールを抽出する手法を提案する．本手法では，加藤らの手法と同様に，コーパス中に頻出する構文構造は正しい構文構造だと考える．本手法の概要は以下のとおりである．

1. 半構造データマイニングの代表的なアルゴリズムである FREQT を用いて，コーパス中に頻出する構文構造を列挙する．
2. FREQT を拡張することにより，頻出する構文構造へ変換できる非頻出の構文構造を列挙する．これは誤った構文構造の候補である．
3. 頻出する構文構造と誤った構文構造の候補とを対にし，後者を前者に変換する誤り訂正ルールを抽出する．

構文木コーパスである Penn Treebank を用いて実験を行い，加藤らの手法では獲得できなかった誤り訂正ルールを獲得できることを確認した．以下に本手法で獲得した誤り訂正ルールの実例を示す．左の構文構造は DT と NN，そして PP で NP を構成しているが，この三つで直接 NP を構成することはできないため，誤った構文構造である．まず DT と NN で NP を構成し，さらにその NP と PP で NP を構成する右の構文構造が，正しい構文構造である．誤り訂正ルールにより，コーパス中に出現する左の構文構造を，右の構文構造に訂正できる．



誤り訂正ルールの例