大規模コーパスの統計量に基づく日本語文章における誤用検知

鈴井　克徳

多くの国・地域では，教育機関で日本語を学ぶ学習者数が引き続き増加傾向にあり，世界各地で日本語学習の需要は拡大しているといえる．また，独学で日本語を学習している学習者等についても，インターネット環境の普及に伴い増加しているとの報告が上がっている．このため，教育機関での日本語の授業以外での自習や独学で日本語の独習を行う学習者を支援する Web サービスやコンピュータソフトウェアの必要性も大きくなってきていると考えられる．そのような Web サービスには，母語でない言語で文章を作成し，それを母語話者に読んでもらう，言語交換というコミュニケーションを媒介する，Lang-8 のような SNS がある．Lang-8 においては，ユーザの投稿頻度から，アクティブユーザの定着が難しいということがいえる．この原因として，母語でない言語で文章を書く労力が，学習者の独習においては大きすぎるということが挙げられる．

日本語学習者の日本語での作文を支援する方法として，本研究では，日本語学習者の文章に含まれる誤用と考えられる箇所を自動で検知する手法を提案する．はじめに，日本語話者による日本語の自然言語文章コーパスを用意し，各文章について，形態素解析と係り受け解析を行った結果のデータベースを構築する．日本語学習者の書いた文章についても，形態素解析と係り受け解析を行い，得られた文章中の係りと受けの組それぞれについて，データベース中の，係りの文節を含む文章数，受けの文節を含む文章数，係りと受けの文節をともに含み，両者に係りと受けの関係がある文章数を調べる．これら 3 種類の文章数を基に，自己相互情報量（PMI）を計算する．この PMI が閾値より小さい係り受けについては，誤用であるものとして検知する．この際，コーパスに出現しにくい固有名詞や特殊な単語を含む文節についても PMI を計算するために，単語の上位下位関係を用いて，分節中に含まれる名詞を上位語に置き換えた文節についても検索する．

実験では，Lang-8 に投稿され，添削された日本語学習者の日本語の文章を用いて，それぞれの文章の不自然さの推定を行った．添削者によって訂正された度合いと，提案手法によって推定された不自然さの度合いとの相関係数を求めた．結果として，両者の値には弱い相関があることが確認されたが，上位語拡張をせずに不自然さを推定した場合の方が，より高い相関が確認された．

（指導教員　若林　啓）