

修 士 論 文 要 旨

開放環境科学 専攻	学籍番号 80325836	氏名 本田 ヒロシ	本田 ヒロシ
(論文題目)			
機械学習によって獲得されたパーサーのメンテナンスに関する研究			
(内容の要旨)			
<p>インターネット上に存在する文書は爆発的に増え続けていると言われて久しい。その一方、一つあたりの WEB サイトに存在する文書の量も非常に増えている。そのため、ユーザがサイト内で必要な文書を探すことは以前よりも困難になってきている。そこで、サイトに情報検索システムや対話システムを設置することによってユーザが容易に情報を見つけられるように工夫がされている。これらの自然言語処理システムでは容易にメンテナンスを行えることが必要である。これは、システムに対するユーザの応答が多様であることと、サイトに存在する文書は時間を経るたびに増加していくことが理由である。また、これらのシステムは汎用である必要はなく、サイト内の範囲で機能すれば十分である。現在、このような自然言語処理システムのためのパーサーの構築が必要になってきている。</p> <p>そこで、本研究では比較的小規模だが、メンテナンスが容易な英語のパーサーを獲得する手法および、獲得した英語のパーサーのメンテナンスを行う手法を提案した。</p> <p>パーサーの獲得には帰納論理プログラミングを用いた。帰納論理プログラミングは学習データが増加すると、それにもなって学習時間も非常に増えてしまう。本研究では学習時間の増加を防ぐため、非終端記号ごとに独立な学習を行った。獲得したパーサーのメンテナンスはパーサーから出力される構文木もしくはその部分木を人間が判断し、帰納論理プログラミングの学習データに追加することによって行った。</p> <p>パーサーを獲得するシステムには 350 文の学習データを与えた。獲得したパーサーに対して 80 文のテストデータを用意し、テストを行った。その内、正しい構文木を一つだけ出力したものは 17 文 (21.3%)、正しい構文木を含む複数の構文木を出力したものは 32 文 (40.0%) であった。残りの 31 文 (38.8%) は誤った構文木のみを出力した、もしくは何も出力しなかったものである。正しい構文木を一つだけ出力したものを除く 30 文を使ってメンテナンスを行った。その結果、テストデータ 50 文のうち 21 文 (42.0%) が正しい構文木を一つだけ出力した。これによって、比較的小規模だが、メンテナンスが容易なパーサーを構築することができることが示せた。</p>			