平成 10 年 2 月 23 日

電子・情報工学専攻	学籍番号	913701
申請者氏名	渥美 清隆	

指導教官氏名	増山 繁 教授 磯田 定宏 教授

論 文 要 旨 (博士)

論文題目	Basic Studies on Natural Language Processing Techniques Related to Error Correction (自然言語処理における誤り検出および訂正の基礎的研究)

(要旨 1,200 字程度)

音声や文字認識ソフトウェアなどによって入力された大量のデータは多くの誤認識文字を含んでいる. 過去の多くの研究によって, これら誤認識文字は減りつつあるが, まだ完全に無くなったわけではない. そして, 誤認識文字を含んだ大量の文書データから自動的に情報を抽出することは不可能である.

n-gram を用いた方法, あるいは n-gram を拡張した方法によって誤り検出・訂正を行う方法が過去において精力的に研究されてきたが, n-gram のみに頼った方法は限界に達しつつある. この限界を打破するために一文内の形態素情報, 構文情報, 意味情報, あるいは談話内の文脈情報を用いた方法が期待されている.

本研究は一文内の構文情報, 意味情報を利用するための方法の提案と, そこから発生するいくつかの問題を解決した. また, 今まで使われてきた n-gram の方法の性質を調査した. これらの研究成果により, n-gram の方法に構文情報や意味情報を加えた新しい方法を開発することが容易になった. 例えば, n-gram の出力結果から構文解析に用いる導出規則の重み付けを変更することによって, n-gram の結果を重視しつつも, 構文的にも正しい出力結果を得ることができる. 以下に具体的な研究成果を挙げる.

まず, 一文内の構文情報, 意味情報を利用した誤り検出・訂正方法の提案として, コンパイラにおけるプログラムの構文的な誤りを発見するためのアルゴリズムを拡張した. 与えられた文脈自由文法に対して誤り生成規則を追加し, 誤りを含んだ入力文に対して, この誤り生成規則の使用回数が最小から最小 $+d$ までの全ての解析木を出力するように拡張した. また, 複数の解析木を出力することで, 誤りが含まれる前の文を意味解析から推定することが可能となった.

提案した誤り訂正手法は非常に大量の解析木を出力する. また, Earley 法と呼ばれる構文解析手法に基づいているため, 速度的にも不満がある. これらの問題を解決するために, 次の 2 点を検討した. (1) 確率文脈自由文法を導入することで大量の解析木を抑制しようとした場合, どの程度 “強力に” 抑制することができるのかを, 理論的な側面と実験的な側面の両面から検討した. その結果, 確率値の大きいごく少数の解析木が確率的に十分な情報を持っていることが分った. また, (2) 解析の高速化として並列計算機上での実行を考え, 並列計算機上で効率良く動作する構文解析法を提案した. この方法は解析テーブルと呼ばれるデータ構造に対して要素を作成する順番を工夫することによって, 入力列長 n に対して, プロセッサ数は $O(n^{4.752})$ 個, 実行時間は $O(\log^2 n)$ で実行可能である.

一方, n-gram を用いた方法における誤り検出の性質の検証では, 比較的よく使われる tri-gram を対象として, 学習用コーパスの種類や量を変更しながら, 再現率と適合率の調査を行なった. この結果, 再現率にはあまり大きな変化はみられなかったが, 適合率はコーパスの種類や量に大きく影響されることが分った. 適合率とコーパスの種類や量の関係を示す要素として網羅率を提案した.