平成１４年度　情報工学コース卒業研究報告要旨

稲垣 研究室	氏　　名	大　　野　　誠　　寛
卒業研究題目	統計的構文解析器を用いた 構文木付音声言語コーパスの構築に関する研究	

大規模係り受けコーパスは，自然言語処理において重要な役割を果たしている．実際，新聞等の大規模言語データに基づく書き言葉の係り受けコーパスは，これまでに広く利用されている．それに比べ，話し言葉の係り受けコーパスは十分に整備されているとは言い難い．係り受けコーパスは，統計的な係り受け解析の統計情報として利用でき，大規模な音声言語係り受けデータを作成し，活用することにより，ロバストで精度の高い音声言語係り受け解析の実現が期待できる．しかしながら，係り受けコーパスの構築には，一般に，形態素情報，文節まとめあげ情報，ならびに，係り受け情報を付与する作業が必要であり，それを人手で行うのは莫大な労力をともなう．

そこで本論文では，統計的構文解析器を用いた大規模音声言語係り受けコーパスの構築手法を提案する．本手法では，まず，形態素・構文情報を付与したデータを自動的に作り上げ，それを人手により修正する．また，統計的な構文解析器は，学習データが増加するほど高精度な解析の実現が期待できることに着目し，本研究では，統計的構文解析と人手修正により構築したコーパスを，さらに規模の大きいコーパスを構築するための統計情報として利用するという増殖的なコーパス構築手法を採用した．構築対象のコーパスデータの全てに係り受け解析を施し，その後一括して人手で修正を行うよりも修正に要するコストが軽減できると予想される．

本手法に基づき音声言語係り受けコーパスを構築するために，ロバストな統計的構文解析器を開発し，それを用いて名古屋大学 CIAIR 車内音声対話コーパスに収録されている 45,053 文節からなる 10,995 ターンのドライバー発話に対し，形態素・構文情報を付与する作業を実施した．人手のみにより作成した 4,277 文節を除く，40,776 文節に対する係り受け解析の正解率は 89.0%であった．これは，統計的構文解析器の使用により，修正するデータ量の 10 倍程度の規模を備えた係り受けデータの構築が可能になることを意味しており，本手法の効果を示している．

学会発表実績等

- 電気関係学会東海支部連合大会　(2002.9)
  “係り受けに基づく話し言葉コーパスの統計的分析”,　p.245
- 言語処理学会第 9 回年次大会　(2003.3 発表予定)
  “統計的構文解析器を用いた音声言語係り受けコーパスの構築”
- 情報処理学会第 65 回全国大会　(2003.3 発表予定)
  “日本語音声対話文の統計的係り受け解析とその評価”