氏名	森（もり） 信介（しんすけ）
学位(専攻分野)	博士（工学）
学位記番号	工博第1691号
学位授与の日付	平成10年3月23日
学位授与の要件	学位規則第4条第1項該当
研究科・専攻	工学研究科電子通信工学専攻
学位論文題目	テキストコーパスからの確率的言語モデルの推定
論文調査委員	（主査）教授 長尾 真　教授 松山隆司　教授 吉田 進

論文内容の要旨

自然言語を確率的現象としてとらえる確率的言語モデルの考え方は古くからあったが，そのモデルの規模は小さく，またあまり大量でない言語データから確率を推定していたために，現実の言語モデルとして不満足なものであった。本研究では，大規模な言語データに対してクロスエントロピーを目的関数とし，これを最小にするという基準を用いて確率正規文法（n-gramモデル）と確率文脈自由文法などを具体的に構成し，文字や形態素，文節などの予測力を向上する方法を示したもので，9章からなっている。

第1章は本論文の目的について述べている。

第2章は確率的言語モデルについて述べ，テストコーパスを参照することなく，そのクロスエントロピーが小さくなるモデルを推定することが有意義であることを明らかにした。

第3章は日本語文について文字単位の確率的言語モデルについて述べている。文字n-gramモデルのパラメータ推定の一般的手法を明らかにし，削除補間法を用いることによって既知文字集合の決定法が自然に導出できることを示した。種々の実験の結果，削除補間における学習コーパスの分割数は10程度がよいなどの結論をえた。

第4章は日本語の形態素単位の確率的言語モデルについて述べている。単語表現とその品詞対を形態素とする基本単位を作り，これを前章の文字の代りとしたモデルを扱った。ここでは未知語文字列の予測も含めた新しいモデルを提案している。また学習コーパスの大きさと先行事象の長さのクロスエントロピーに対する影響を実験的に示し，モデルの実用化のための指針を与えた。

第5章は日本語の形態素クラスタリングについて述べている。n-gramモデルを形態素について行うとき，クラスタリングの手法を導入して形態素をクラス分けし，そのクラスを用いてn-gramモデルを作るクラスn-gramモデルを作った。クラスタリングにはテストコーパスの平均クロスエントロピーを目的関数として用いた。その結果クラスn-gramモデルが形態素n-gramよりも，少なくともn=2に対してはすぐれていることを示した。

第6章は日本語形態素解析への応用について述べている。確率的形態素解析の方法を説明し，未知語モデルに外部辞書を付加する方法の提案を行った。実験の結果この方法はこれまでの形態素解析の方法よりもよい精度を与えることを示した。

第7章は日本語の文節を単位とした確率的言語モデルについて述べている。文節を予測単位とし，係り受けという構造を内包するモデルで，これは確率文脈自由文法に属するものである。このモデルのパラメータをこれまでと同じ手法で求め，これをこれまでの各種モデルと比較した結果，予測力という点でよりよいモデルとなっていることを明らかにした。

第8章は係り受けモデルを用いた日本語の確率的構文解析の方法について述べている。その解探索にはCKY法を確率文脈自由文法に拡張したアルゴリズムを用いた。その結果，EDRコーパスにたいしては，80.48%という優れた解析精度を達成している。このモデルの利点は未知語の品詞推定や形態素解析を同時に実行できることである。

第9章は本論文の結論である。

論文審査の結果の要旨

自然言語を確率的現象としてとらえる確率的言語モデルにおいて，その予測単位を文字や，形態素（単語），形態素クラス（品詞），文節などに取り，これらをクロスエントロピー最小という基準によって大規模言語データ（コーパス）から得られる確率的パラメータによって構成し，日本語文の形態素解析や構文解析に応用し，精度の高い解析結果が得られることを示したもので，得られた成果は以下の通りである。

(1) 削除補間法を応用することで定義できる学習コーパス内での平均クロスエントロピーを下げることを目標として，モデルの可変部分のパラメータを決定するという確率的言語モデルの構成法についての新しい手法を提案した。

(2) 形態素クラスタリングという新しい手法を考案し，形態素クラス（品詞）を単位としたクラス n-gram モデルを提案し，予測力の有意な向上が得られることを示した。

(3) 上記のクラス n-gram モデルを用いた確率的形態素解析の方法を示し，未知語モデルに外部辞書を付加することによって優れた形態素解析の精度が得られることを示した。

(4) 文節を単位とする係り受けを確率的文脈自由文法でモデル化し，構文を考慮に入れた確率的言語モデルを提案するとともに，これの改善として形態素クラスタリングの結果を用いたモデルを作り，形態素の予測力の有意な向上があることを示した。またこれを用いることによって構文解析の精度も向上することを示した。また，この方法は未知語の品詞推定や形態素解析を同時に精度高く行えるという利点を持つことを示した。

これらの成果は日本語文の確率的モデルとして十分に実用に供することができるものであって，学術上，実際上，寄与するところが少なくない。よって，本論文は博士（工学）の学位論文として価値あるものと認める。

また平成10年１月16日，論文内容とそれに関連した事項について試問を行った結果，合格と認めた。