論文内容の要旨

博士論文題目　Distributional Approaches to Natural Language Processing
(自然言語処理への分布的アプローチ)

氏名　　持橋大地

(論文内容の要旨)

自然言語は単語，句などの基礎的な言語的要素を構造化することにより様々な意味を表現する複雑な組織体である．したがって，その理解のためには構造化の方法と同様に，基礎的要素の持つ意味についても適切な知見の下に取り扱いを行う必要がある．

統計的自然言語処理においては，構文解析や依存構造解析など構造的要素の統計的モデル化が進んでいるが，その基礎となる単語などの統計的モデル化は近年始まったばかりであり，多くの恣意的な仮定を残している．本論文では，基礎的要素の，コーパスにおける分布的情報から導かれる統計的モデル化に焦点を当て，コーパス上にみられる静的な単位と動的な単位に着目して，それぞれについてこれまで行われてきた恣意的仮定を取り除くことで，分布的情報をより柔軟に取り扱うことを目的とする．

静的単位とは文書，パラグラフ，文など，ある目的のために事前に与えられている単位を意味する．これらの間の意味的な距離を測ることが多くの自然言語処理の基礎をなしているが，カーネル法が使用できない場合は，通常，このためにtf.idf等で重みづけされた素性ベクトル間のユークリッド距離が用いられてきた．しかしながら，この距離は重みづけに恣意的要素を持つこと，また素性間の相関を表現できないという欠点を持っている．本論文ではこの問題に対し，データ中であらかじめ意味的に似ているとされるクラスタ構造の歪みを最小化する最適な計量距離を導入し，前記問題を同時に解決した．同様の意図を持つ先行研究と異なり，この方法は繰り返し最適化を必要とせず，すべてのデータを用いて最適な計量を解析的に決定することができる．同義文検索，文書検索およびベクトルデータのクラスタリングタスクにおいて，従来の重みつきユークリッド距離に対する優位性を検証した．

動的単位とは上のような所与の単位と異なり，統計的分析を行うことで，その内部での均質性が同定される，より精密な単位を意味する．近年，文書を単位とした統計モデルがいくつか提案され，文書中で生起する単語について混合モデルの観点から統計的に取り扱うことができるようになっている．しかしながら，そこでは文書内の意味的均質性が仮定されているために，充分長い文書には適用することができず，また文脈言語モデルなどへの適用において，仮想的な文書として表現された履歴全てを文脈として使用してしまい，適切な予測が行えないという問題を持っている．

これに対し，本論文では文脈の変化を隠れた確率過程としてとらえ，Mean Shift Model と呼ばれる統計モデルを自然言語について拡張することで，隠れた時系列としての文脈を観測単語列からベイズ的に逐次推定する方法を提案した．このモデル

は本質的に非線形な隠れマルコフモデルであり，従来の離散隠れマルコフモデルやカルマンフィルタ等は推定に使用することができないが，任意の非線形時系列を追跡できる逐次モンテカルロ法を導入することで推定を行うことができる．この方法は意味的な変化点を確率的に検出し，そこからの文脈を用いることで最適な予測を得る方法であり，前記の動的単位を時系列的に同定するものとなっている．標準的なBNCコーパスでの実験において，単純な履歴を用いる従来のベイズ文脈モデルに対して常に優れた予測を示した．

静的単位および動的単位に対して分布的情報を適切に扱うこれらの方法により，自然言語処理において，単語など基礎的要素のより柔軟な統計的取り扱いが可能となった．

氏　名	持橋大地

（論文審査結果の要旨）

平成17年1月20日に開催した公聴会の結果を参考に平成17年2月17日に本博士論文の審査を行った．以下のとおり，本博士論文は，提案者が独立した研究者として，研究活動を続けていくための十分な素養を備えていることを示すものと認める．

持橋大地は，本博士論文において，言語の分布情報に基づく統計的言語モデルの研究を行い，静的および動的な視点から以下のような手法を提案し，その有効性を示した．

1. 言語の基本単位である単語および単語から構成される文や文書の意味的近さを測るために，意味的に似ているとされるクラスタ構造の歪みを最小化する最適な計量距離を導入し，恣意的要素の少ない最適な距離計量の推定法を提案した．この方法による距離計量は最適値を解析的に決定できるという利点も持っている．事例に基づく機械翻訳のための同義文の検索，文書検索，および，文書クラスタリングの実験に提案した意味距離を用いた実験を行い，従来手法に対する優位性を示した．

2. 文書に生起する単語を予測する統計的言語言語モデルは，局所的な情報にのみ依存するか，あるいは文書全体といった大局的な情報に基づくことが多く，話題の変化に適切に追従するモデルの提案はあまりなされていなかった．本論文では，文脈の変化を隠れた確率過程としてとらえ，Mean Shift Model と呼ばれる統計モデルを自然言語に対して拡張することによって，隠れた時系列としての文脈を観測単語列からベイズ的に逐次推定する方法を提案した．この方法は意味的な変化点を確率的に検出し，そこからの文脈を用いることで最適な予測を得る方法といえる．BNC コーパスを用した様々な長さを持つ文書データに対する単語予測実験を行い，単純な履歴を用いる従来のモデルに対して優れた予測が可能であることを示した．

言語の分布情報を適切に扱う手法を提案した本研究は，独創性が高く，しかも実用的であり，自然言語処理の分野において高い貢献があると評価する．

よって，本論文は，博士（理学）の学位論文として価値あるものと認める．