論文内容の要旨

博士論文題目　Domain Adaptation of Statistical Word Segmentation System
(統計的単語分割の分野適応手法)

氏名　　坪井 祐太

(論文内容の要旨)

分かち書きされていない日本語や中国語では文の単語への分割は自明でない．これらの言語の単語分割問題においては統計的な手法が適用され，その有効性が示されている．しかし，実際の応用では単語分割を学習したデータと異なる分野への適用時に語彙や文体の違いによる性能の低下が常に課題となっている．そこで学位本論文では，単語分割器の分野適応時に有効な以下の2つの手法を提案する．

1点目は，文の一部にのみ単語境界情報を付与する（部分的アノテーション）手法である．文中の重要と思われる部分のみに集中できることにより，新しい分野の学習データの作成が効率的になる．本論文では文中に部分的にアノテーションが付与されたデータを用いて条件付確率場 (CRF) を学習する手法を提案する．CRF は単語分割問題に適した統計モデルであることが知られているが，既存の CRF の学習法では文全体がアノテーションされたデータを想定していた．そこで，周辺尤度を目的関数にすることで部分的アノテーションを用いて CRF を学習する方法を提案する．

2点目は，適応先分野での性能を最大化するような重要度重みを適応元データに付与する手法である．本論文では学習データとテストデータの入力の密度比をサンプルから直接推定する方法を提案する．提案手法の計算量はテストサンプル数とはほぼ独立なため，提案手法はアノテーションされていない適応先データが大量に入手可能である単語分割問題に適している．

本論文では，上記の提案手法により統計的な単語分割器の適応先での性能向上を計算機実験により確認した．

氏 名	坪井 祐太

（論文審査結果の要旨）

平成20年8月4日に開催した公聴会の結果を参考に平成21年2月19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．

坪井 祐太は，本博士論文において，日本語や中国語のように文中の単語を明示的に分割しない言語の単語分割を統計的に行う際に，学習データとは異なる分野のデータに対する解析精度を効率よく向上させるための分野適応手法手法を提案した．具体的には，次の2つの手法を提案し，日本語の単語分割システムの性能評価を行うことにより，提案手法の有効性を検証している．

1. 学習を行った分野とは異なる分野のテキストは，新規の専門用語や分野固有の言い回しをもつかも知れないが，機能語や一般の動詞や名詞に関する用法が異なるわけではない．適応すべき分野の学習データに対して完全な単語分割を施すのではなく，分野独特の専門用語等の周辺に対する部分的なアノテーション作業のみを課すことは，複雑な言語現象の知識を必要とせず，アノテーション作業の大きな軽減につながる．本論文では，部分的なアノテーションを施されたデータに対しても条件付確率場の学習が可能になるよう，学習法の拡張手法を提案した．

2. 新しい適応分野では，学習に用いた分野のデータと比較して，出現する語や言語現象の出現頻度等の統計量が異なる．本論文では，学習データと，適用先の分野のデータとの密度比を大量のサンプルデータから直接推定する効率的な手法を提案し，適応先の分野のアノテーションされた学習データがない場合にも単語分割精度を向上させる手法を提案した．

このように，本論文では，学習データとは異なる分野のテキストの単語分割を統計的学習によって行うための半教師付きおよび教師なしの分野適応手法を提案した．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（工学）の学位論文として価値あるものと認める．