論文内容の要旨

博士論文題目　Machine Learning Approaches to Rhetorical Parsing and Open-Domain Text Summarization
(修辞構造解析とテキスト自動要約：機械学習による接近)

氏名　野本忠司

(論文内容の要旨)

本論文では，日本語の修辞構造解析とテキスト自動要約について機械学習を応用した幾つかの手法について提案と検討を行う．

修辞構造解析では，主として文の間の修辞的依存関係と修辞タイプの同定問題をとりあげる．ともに日経新聞を記事を用いて，人手で依存関係，タイプをマークアップしたデータをベンチマークデータとして，確率的決定木を導入し，依存関係，タイプ同定の予測性能を調べる．また，ひとつの拡張として，決定木をランダム生成したコミッティベースの能動学習を提案する．

テキスト要約は，その目標として要約とは直接関係のない外的タスクでの有効性を目指す立場と内的タスクでの有効性 (つまり、人間なみの要約の生成) を目指す立場がある．本論文では，それぞれの立場での要約タスクを教師なしおよび教師あり学習等，異った機械学習の観点でアプローチする。

まず，外的なタスクとして本稿では文書検索を考える．これは，文書検索での有効性が要約のよしあしを決めるという立場であり，要約とは，文章としての自然さより，元文書の主要な情報を保持していればよいという考えかたが背後にある．本稿ではMDLで若干拡張したクラスタリングベースの要約手法を提案し，BMIR-J2と呼ばれる文書検索のベンチマークデータで手法の有効性を検証する．

一方，内的タスクとして，本稿では人手による重要文判定データを収集し，判定データをモデリングすることを考える．具体的には，確率的決定木をベースにした判定データを直接学習する手法と，判定データを全く参照しないクラスタリングに基づく手法とを比較し，両者の精度と判定者間の一致度との関係を見る．

さらに，本稿では，被験者判定の分布がドメインに固有であることに着目して、分布自体を直接統計的にモデリングして要約の生成をおこなうことを考える．このアプローチでは，分布情報のみを使って要約を生成するところに特徴がある．ベンチマークデータとして人手による重要文判定データを用いて，kNN, Naive Bayes, 決定木等，知識集約型学習手法とパフォーマンスの比較をおこなう．最後に，このアプローチが人手要約の大きな特徴である一致性と不一致性を自然に表現できることを示す．

氏 名	野本忠司

（論文審査結果の要旨）

平成16年10月26日に開催した公聴会の結果を参考に平成16年11月26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．

野本忠司は，本博士論文において，機械学習を用いた文書の修辞構造解析と文書要約に関する研究を行い，種々の有用な知見を得るとともに，実用につながる有効な結果を報告している．特に，次のような手法の提案と評価実験を行っている．

1. 確率的決定木を用いて，人手で依存関係や修辞タイプを付与したデータから修辞構造解析規則を学習する方法を提案した．

2. 少ない学習データでの学習を目指して，サンプリングの視点からコミッティベースの能動学習法を，学習器の視点から最小記述長原理やブースティングと決定木学習の組合せによる手法を提案し，比較検討した．

3. 機械学習を用いた領域非限定の文書要約の種々の手法について提案と考察を行った．特に，教師付き，および，教師なしという視点から，手法と評価法を提案した．

4. 文書のトピックの多様性に注目した要約手法として，文のクラスタリングに基づく重要文の選択法を，K-mean クラスタリングと記述長最小原理の組合せという手法により提案した．

5. 文書要約の評価手法として，人手によって選択した重要文による手法だけでなく，要約文書による文書検索の性能を客観的に測るという新たな視点を提案し，その有効性を検証した．

6. 特定の文書ごとに重要文の出現する位置の分布が異なることを利用し，重要文の分布を利用した要約文のための文抽出法を提案した．また，文書のないように深く入り込まないこの手法が，他手法に比較して遜色ない性能を達成することを示した．

その他にも機械学習に基づく様々な文書要約手法を提案し，比較検討したた本研究は、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。