博 士 学 位 論 文 審 査 要 旨

学位申請者氏名	高野　海斗
論　文　題　目	金融テキストを対象とした有益情報抽出に関する研究
審査委員（職名・氏名・印）	
主　　査	准教授　　酒井　浩之
審査委員	教　授　　中野　有紀子
	教　授　　世木　寛之
	教　授　　増山　繁
論文審査結果（合　否）	合　格

論文審査の要旨

金融業界では人工知能分野の手法や技術を金融市場における様々な場面に応用することが期待されており，膨大な金融情報を分析し投資判断を支援する技術に注目が集まっている．その中でも特に，金融テキストから投資判断において有益な情報を抽出し，抽出された情報と市場変動の関係性を発見し，市場分析に応用する研究は金融テキストマイニングと呼ばれている．投資判断において有益な情報とは，具体的には株価に影響を与えるような情報であり，営業利益のような業績結果だけでなく，そのような業績になった要因，さらには，役員人事，配当の実施の有無など，様々な情報が該当する．

このような背景の下，本論文では，金融テキストを対象として，投資判断において有益な情報を抽出する手法を提案する．第1章では，本研究の目的，自然言語処理分野への学問的な貢献，関連研究についてまとめている．第2章，第3章，第4章が本学位論文の中核をなし，申請者が筆頭著者となる3編の論文を再編成した内容となっている．以下に，各章の内容を簡単に説明する．

第2章，第3章では，金融テキストである株主招集通知を対象にした，投資判断において有益な情報の抽出に関する研究について述べる．株主招集通知とは，株主総会の開催前に株主へ送付される文書であり，その記載内容としては，株主総会で議論される決議事項，大株主情報，役員情報などである．その中には，配当の実施，役員人事など，株価に影響を与える可能性のある情報が多く記載されているが，ページ数が百ページを超えるものも珍しくない．したがって，投資判断に有益な情報が多く記載されているが，発行時期が株主総会開催時期に集中することもあり，資産運用業務で企業分析を行う部署では，株主招集通知から投資判断に有益な情報を確認する作業に膨大な労力を割いている．このような課題を解決するために，本研究では投資判断に有用な情報が何ページから何ページに記載されているかの推定を，ある程度まとまった文の集合であるページ単位で実現する方法論の検討を行う．本研究により，確認したい情報が何ページから何ページまで記載されているのかを自動で推定することが可能となり，大幅な作業の効率化が実現可能になった．

本研究が従来の研究と大きく異なる点は，情報抽出のために文単位で区切り位置を推定するのではなく，ページ単位で区切り位置を推定する点と，モデルを学習するための学習データを自動生成している点が挙げられる．ページ単位で抽出を行うことにより，文単位での抽出よりも学習データの自動生成は容易になるが，ページの途中で次の記載内容が始まるといった独自の問題がいくつか存在する．また，自動生成

論文審査の要旨（続）

で得られる学習データは，人手で作成した場合に比べ，精度やデータの偏り（バイアス）の問題が存在するため，それらの特性を考慮した上で学習させるモデルを選択する必要がある．そこで本研究では，人手で作成された学習データを用いた上で，該当ページに記載されている決議事項である議案の分類が可能であるかどうかの検討を行い，良好な精度で分類が可能であることを示した上で，ページ単位での分類や抽出にどのような問題があるのかの考察を行った．さらに，決議事項である議案だけでなく，大株主情報や役員情報などの情報が何ページから何ページに記載してあるかの推定を自動生成した学習データを用いて行うことにより，学習データの自動生成によってどのような問題が生じるのかを議論した上で，それらの問題を軽減できるモデルの検討を行った．本手法の学習データ自動生成により，人手による学習データ作成では生成することができない大量の学習データの生成が可能となったが，その反面，学習データに偏りが生じていることが明らかになった．そして，この偏りにより，本研究で扱うページに対しての分類を行う系列ラベリング問題において，CRF などの従来研究で良好な結果が得られる手法が必ずしも最良の結果が得られる保証がない．そこで本研究では，いくつかの従来モデルとの検討も行った上で，本手法による BiLSTM モデルがミクロ F1 値 0.970 と最も良好な結果であることを示し，従来の研究で有効とされている CRF 層などの追加が，自動生成した偏りのある学習データを用いる場合には過学習の原因になっていることを示した．

第4章では，同じく金融テキストである有価証券報告書を対象にした，投資判断において有益な情報の抽出に関する研究について述べる．有価証券報告書は，事業年度ごとに作成する企業内容の外部への開示資料であり，企業の概況，事業の状況，設備の状況など多くの内容が記載されており，そのテキスト情報は膨大である．そこで，本論文では有価証券報告書の 2 章「事業の状況」を対象に，投資判断において特に有益な情報である，企業の業績に関する要因について書かれた文（業績要因文）と，どれだけの売上高や経常利益だったのかについて書かれた文（業績結果文）の抽出を行い，抽出した業績要因文，業績結果文が，対象企業のどの事業セグメントに関するものであるかを自動付与する手法について述べる．上記を実現することにより，例えば，自動付与した事業セグメントの情報から，業績要因文に対応する業績結果文が得られるため，セグメント別の業績の要因だけでなく，売上高からセグメントごとの事業規模や，経常利益の前年度比からセグメント別の事業の成長度合いを把握することも可能になる．

業績要因文の抽出は，決算短信を対象に学習データを自動生成し，分類モデルを学習させることで抽出が可能であることが先行研究によって示されているため，その従来手法を本論文で対象としている有価証券報告書に適用し，業績結果文の抽出はルールベースで行った．そして，有価証券報告書の特徴を利用することで，事業セグメントもルールベースで抽出し，抽出した業績要因文と業績結果文が，どの事業セグメントに対する記述であるかを，文の出現順序に着目することで自動付与した．評価実験の結果，事業セグメントの付与が正しく，業績要因文判定が正しいものの適合率は 0.693，再現率 0.725 であり，事業セグメントの付与が正しく，業績結果文判定が正しいものの適合率は 0.788，再現率 0.911 であり，良好な適合率再現率を達成している．

本論文は，金融テキストである株主招集通知を対象に，ある程度まとまった文の集合であるページ単位での有益情報の抽出を自動生成した学習データを用いて行うための手法の提案を行っており，これらは有用性の高いものである．また，金融テキストである有価証券報告書を対象に，文単位での有益情報の抽出として事業セグメントが付与された業績要因文と業績結果文を抽出する手法の提案を行っている．これにより，事業セグメント，業績要因，業績結果の3つが紐づいたデータの抽出が可能となり，より発展的な分析が可能となることが期待できる．よって本論文の内容は，自然言語処理への学問的な貢献，金融テキストマイニングという工学的応用に貢献するものであり，博士（理工学）の学位にふさわしいものである．

（以　上）