成蹊大学大学院　学位論文要旨
2021 年 1 月　高野 海斗

金融テキストを対象とした有益情報抽出に関する研究

本論文は，企業が公開している金融テキストを対象に，投資判断に有益な情報を抽出することを目的としている．金融テキストから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おり，本研究はこの研究の一環である．本論文では，金融テキストとして株主招集通知と有価証券報告書を対象とし，それらから有益な情報を抽出するための手法について述べる．

本論文の自然言語処理分野への学問的な貢献は以下の２つである． １つは，金融テキストにおけるページ単位での情報抽出を，自動生成した学習データを用いて行うための方法論の提案している点である．本論文の 3 章で提案する，株主招集通知から自動生成した学習データを用いたページ単位での情報抽出の方法は，株主招集通知に限らず，多くのデータに対して，応用が可能な汎用性の高いものである．もう１つは，有価証券報告書から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ことが可能になることである．これまで，業績要因文の抽出を行う研究はあったが，業績要因文に対して，事業セグメントや業績結果文を紐づけることで，様々な応用分析が可能になる．

本論文の目的は，大量の人手で作成した学習データがあれば，現在の技術で解決する可能性が高い．まず 2 章では，人手で作成した学習データを用いて分類器を学習し，分類対象となるページをルールベースで絞った上で，ページ単位での分類が可能であることを示す．しかし，機械学習手法を利用するためには，多くの学習データが必要であるが，学習データの作成には多大な労力や高い専門知識が必要である．また，一時的に学習データを人手で作成したとしても，データの特徴は時間経過に伴い変化する可能性が高く，その都度，学習データの作成を繰り返すのは現実的に不可能である．そこで本論文の 3 章と 4 章では，学習データを少ない労力で大量に作成する方法を提案する．これにより，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るが，人手が介入しないため，学習データに偏りが生じることが評価実験によって明らかになった．したがって，人手で作成した学習データを用いるときに良好な結果が得られるモデルが必ずしもうまくいく保証がないため，いくつかの従来モデルの検討も行い，提案手法の有効性を示す．