

論文審査の要旨および学識確認結果

報告番号	甲 第 号	氏 名	Zhang Yejian
論文審査担当者：	主査	慶應義塾大学教授	博士（工学） 高田 眞吾
	副査	慶應義塾大学教授	工学博士 斎藤 博昭
		慶應義塾大学教授	博士（情報学） 杉浦 孔明
		慶應義塾大学教授	博士（工学） 矢向 高弘
(論文審査の要旨)			
学士（工学）、修士（工学）Zhang Yejian 君提出の学位請求論文は「Cost-Efficient Text Classification without Manually Labeled Data（手動ラベルなしによる効率的なテキスト分類）」と題し、全5章から構成されている。 テキスト分類に関する技術は、レビューの感情分析や有害コメントのフィルタリングなど、様々なところで応用されている。多くの手法は機械学習に基づいているが、それらは大量の訓練データを必要としている。訓練データには、個々のデータがどのように分類されるかの情報を含んでいるが、そのような情報は手作業で付加されており、そのコストは無視できないほど高い。さらに、実行環境も高CPU・高メモリであることを前提にしている問題もある。その結果、分類手法そのものは精度がよくても、実用性の観点から難点がある。そこで、本論文では、これらの問題をソフトウェア工学の視点から検討し、二つの分類システムを提案、実装、そして評価を行っている。訓練データの手作業に関わるコストを下げると共に、手元に高CPU・高メモリの実行環境がなくても、高い精度で実行できることを実現している。 第1章は序論であり、本研究の背景、目的、および論文の構成について述べている。 第2章では、関連研究を示している。本研究で用いている基礎研究について述べた後、本研究で提案する二つのシステムを、既存研究とどういう関係にあるのかを明確にしている。 第3章では、少量のラベルを用いて、ゲームレビューをコストの観点から効率よく分類するシステムを提案している。提案システムでは、まず、自己教師あり学習を通して、ゲームレビュードメインに特化した言語モデルを学習する。次に、教師なし学習と能動学習を組み合わせることにより、少量の手動ラベル付けを行う。評価の結果、提案システムは3クラス分類タスクにおいて100個のラベル付きデータのみで88.8%の分類精度を得た。また、従来手法に比べて大幅にコストを削減できている。 第4章では、能動学習の中で、大規模言語モデルをオラクルとして用いるシステムを提案している。4種類のクエリ戦略に基づき、ラベルのないデータプールからデータを選択し、そのデータをオラクルがラベル付けする。選択するデータは、不確定性や多様性など、何らかの観点で今までにないデータである。提案システムは、感情分析、ニュース分類、有害コメント抽出といった三種類の分類タスクに対して評価している。その結果、従来手法に比べて、精度はほとんど変化せず、訓練などのためにかかるコストはおおよそ6%ですみ、実用性を示している。 第5章は結論であり、本研究で得られた成果を総括し、今後の研究の展望について述べている。 以上要するに、本論文はソフトウェア工学の観点から機械学習における訓練コストの問題に着目し、それを解決するためのシステムを二つ提案・構築した点で工学上寄与するところが少なくない。また、これらの成果は著者が自立して研究活動を行うために必要な高度な研究能力、並びにその基礎となる豊かな学識を有することを示したと言える。 よって、本論文の著者は博士(工学)の学位を受ける資格があるものと認める。			
学識確認結果	学位請求論文を中心にして関連学術について上記審査会委員で試問を行い、当該学術に関し広く深い学識を有することを確認した。 また、語学（英語）についても十分な学力を有することを確認した。		