氏　　名	Apichai Chan-Udom
博士の専攻分野の名称	博士（工学）
学 位 記 番 号	医工博甲４５３号
学 位 授 与 年 月 日	平成３１年3月２０日
学 位 授 与 の 要 件	学位規則第４条第１項該当
専　　攻　　名	情報機能システム工学専攻
学 位 論 文 題 目	A Study on Knowledge Acquisition from Social Media Data and Its Application（ソーシャルメディアからの知識獲得と その応用に関する研究）
論 文 審 査 委 員	主査　教　授　鈴木　良弥
	教　授　宗久　知男
	教　授　福本　文代
	准教授　渡辺　喜道
	准教授　丹沢　勉
	准教授　西崎　博光

学位論文内容の要旨

ツイッターに代表されるソーシャルネットワークサービスは全世界で非常に多くの利用者がおり，その利用者が毎日膨大な量の文書を投稿している．投稿された文書には最新の情報が含まれており，その新しい情報から多くの知識を獲得する研究が多数行われている．本論文はツイッターに代表されるソーシャルメディアサービスへの投稿データから知識を獲得しようという研究である．不特定多数のユーザが即時性を重視して発信するため不均質なテキストデータから語義の曖昧性解消，新語への意味付与，不均質なショートテキストの分類に対して，技術的な困難さについて言及し，その解決策を提案し，一定の成果を報告した．

具体的には，

１．「英語ツイートデータを用いた科学技術用語の意味同定」（博士論文２章，３章に対応）

２．「英語ツイートデータ中の科学技術用語の抽出とその意味検出」（博士論文２章，３章に対応）

３．「同一イベントに関するタイ語ツイートデータと日本語ツイートデータの分類とツイート中からのアイデアの抽出」（博士論文４章に対応）

について述べている．

１．「英語ツイートデータを用いた科学技術用語の意味同定」では，Pacling2017 (the 15th International Conference of the Pacific Association for Computational Linguistics)への投稿論文“Identification of Word Sense in Twitter Data Based on WordNet Glosses”（2017 年 8 月）に関する研究をまとめている．プリンストン大学のジョージ・ミラーらが構築している語彙データベースである WordNet の語義文を利用し，ツイート内に出現する技術用語が WordNet 内のどの語義に対応するかを判定する手法を提案している．単語類似度を求めるために 6 種類の尺度（Hellinger Distance(TF-IDF), Cosine 類似度(Binary), Cosine 類似度(TF-IDF), Hellinger Distance(term frequency), Jaccard Coefficient, Cosine 類似度(term frequency)を用いて実験を行い，Hellinger distance (TF-IDF)を利用した時に語彙の意味同定精度が最も良いことを示した．

２．「英語ツイートデータ中の科学技術用語の抽出とその意味検出」では，LTC’17 (The 8th Language and Technology Conference)への投稿論文 “Detection of new words and their senses in Twitter data using Wikipedia” (2017 年 11 月)に関する研究をまとめている．１．の「英語ツイートデータを用いた科学技術用語の意味同定」では対応できなかった WordNet に登録されていない新語の意味検出を行っている．WordNet に登録されていない 66 種類の単語を用いて実験を行い，それらの単語を Wikipedia の適切なページに対応付けることができた．

３．「同一イベントに関するタイ語ツイートデータと日本語ツイートデータの分類とツイート中からのアイデアの抽出」では，CISAI2018 (The 2018 5th International Conference on Systems and Informatics）への投稿論文 “Classification of Thai Tweets: Mining Treasures from Tweet Heap” （2018 年 11 月），CICLing2019 (20th International Conference on Computational Linguistics and Intelligent Text Processing)への投稿論文“Classifying Short Text in Social Media for Extracting Valuable Ideas” (2019 年 4 月)に関する研究をまとめている．2018 年 6 月～7 月にタイ王国の洞窟で起きた少年らの遭難とその捜索，救出に対するツイートは世界各国から投稿された．その中でタイ語ツイートと日本語ツイートのデータを利用してツイートを 4 種類（救出方法の提案，感情表現，報道，その他）に分類した．ツイート利用の特徴から感情表現（「がんばれ！」，「無事を祈っています」など）のツイートが多いが，その中に少数の「救出方法の提案」が含まれている．その「救出方法の提案」を抽出するためにツイートを 4 種類（救出方法の提案，感情表現，報道，その他）に分類している．ツイートの分類には多くの研究があるが，ほとんどが 1 つのツイートに 1 つのラベルを付けている．しかし，実際のツイートは複数の内容が含まれている．例えば報道内容を引用した後，自分の感情を表現するツイートが多い．そのためこの研究ではマルチラベルの分類を行っている．ツイートの分類のために 5 種類の分類器（Support Vector Machines (Linear), Naïve Bayes (Gaussian), Naïve Bayes

(Bernoulli), Stochastic Gradient Descent, Passive Aggressive Classifier）を用い実験を行い, 実験に用いた 5 種類の分類器の中では Support Vector Machines (Linear)を用いたときの分類精度が一番良いことを示した.

論文審査結果の要旨

ツイッターに代表されるソーシャルネットワークサービスは全世界で非常に多くの利用者がおり，その利用者が毎日膨大な量の文章を投稿している．その投稿には最新の情報が含まれており，その新しい情報から多くの知識を獲得する研究は社会的・学術的意義が大きい.

本論文ではツイートデータから知識獲得の研究について述べている．具体的には
研究１．「英語ツイートデータを用いた科学技術用語の意味同定」
研究２．「英語ツイートデータ中の科学技術用語の抽出とその意味検出」
研究３．「同一イベントに関するタイ語ツイートデータと日本語ツイートデータの分類とツイート中からのアイデアの抽出」
について述べている.

研究１．「英語ツイートデータを用いた科学技術用語の意味同定」ではプリンストン大学のジョージ･ミラーらが構築している語彙データベースである WordNet の語義文を利用して，ツイート内に出現する技術用語が WordNet 内のどの語義に対応するかを判定する手法を提案している．語義の曖昧性解消の研究は古くから研究が行われているが，1990 年代の研究から最近の研究の調査を行い，従来手法の技術的課題を見つけ，その課題を解決するための手法を提案している．本研究の技術的貢献は語の曖昧性解消語彙の類似度の計算に word2vec で作成した単語の分散表現と WordNet の語義文を用いて多義語の意味の同定の精度を高めたことである．意味同定の結果は WordNet の第一語義を正解としたときとの比較を行っている．また語義の同定に失敗した場合の分析・考察を行っている.

研究２．「英語ツイートデータ中の科学技術用語の抽出とその意味検出」では１．の「英語ツイートデータを用いた科学技術用語の意味同定」では対応できなかった WordNet に登録されていない新語の意味同定を行っている．WordNet には非常に多くの語彙が登録されているが，最新の技術用語などはあまり登録されていない．また新しい語は日々増加していくため，人手で WordNet などの辞書に登録することは非常にコストがかかる．そこで WordNet に登録されていない語彙に対して Wikipedia のデータを用いて単語の意味を同定する方法を提案している．Wikipedia には同じ単語に対して複数のページが存在するが，Wikipedia の各ページのアブストラクトを利用することで, ツイート内で使用されている意味を特定している.

研究３．「同一イベントに関するタイ語ツイートデータと日本語ツイートデータの分類とツイート中からのアイデアの抽出」では，2018 年 6 月～7 月にタイ王国の洞窟で起きた少年らの遭難とその捜索，救出に対するツイートは世界各国から投稿されたが，その中でタイ語ツイートと日本語ツイートのデータを利用してツイートを４種類（救出方法の提案，感情表現，報道，その他）に分類した．ツイート利用の特徴から感情表現（がんばれ，無事を祈っていますなど）のツイートが多いが，その中に少数の「救出方法の提案」が含まれている．その「救出方法の提案」を抽出するために４種類に分類している．
ツイートの分類には多くの研究があるが，ほとんどが１つのツイートに１つのラベルを付けている．しかし，実際のツイートは複数の内容が含まれている場合がある．例えば報道内容を引用した後，自分の感情を表現するツイートが多い．そのためこの研究ではマルチラベルの分類を行っている．分類結果を得た後，コンフュージョンマトリックスにより，どの分類がどの分類に間違いやすいかを分析し，分類間違いしやすいツイートについても言及している．また，分類結果をもとにツイート中で提案されている解決策を抽出できることを示している．
以上より，本論文は不特定多数のユーザが即時性を重視して発信するため不均質なテキストデータから最新の情報を抽出する点に重要性がある．また，その困難な課題に対して，技術的な困難さについて言及し，その解決策を提案し，一定の成果を報告している．

博士論文最終試験は 2019 年 2 月 6 日に行った．質疑応答では，上記の３つの研究の関連性に関する質問，採用した手法に関する質問，分類結果の精度に関する質問に対して，回答を行った．
各研究の関連性に関しては，どの研究も不特定多数のユーザが発信する短く不均質なテキストデータからの知識獲得であり，特に研究１「英語ツイートデータを用いた科学技術用語の意味同定」の課題として，WordNet に登録されていない単語がツイート内に多数出現するが，研究１の枠組みではその新語の意味を同定することが出来ないため，研究２「英語ツイートデータ中の科学技術用語の抽出とその意味検出」で Wikipedia の情報を利用して新語の意味の検出を行ったと回答している．
採用した手法を用いた理由に関しては，提案した手法で用いた技術は最新の手法ではないが，その技術を組み合わせることによって精度を向上させることが出来た．今後の研究として，是非深層学習を用いた最新の手法も試してみたいと回答した．
実験で得られた分類精度の評価についての質問については，同じデータで実験をしている研究がないため明言できないが，複数の分類器を使って分類結果を比較していることを説明した．

博士論文で取り組んだ課題の重要性，提案した手法の有用性を示すための実験と考察，

最終試験でのプレゼンテーションの構成，質疑応答の的確さを鑑み，本論文は，博士論文としてふさわしい研究であると判断した.