様式3

論 文 内 容 の 要 旨

氏 名 （ 中村 達哉 ）	
論文題名	Research on Cross-lingual Analysis of Social Media using Wikipedia (Wikipediaを用いたソーシャルメディアの言語横断解析に関する研究)

論文内容の要旨

Social media has grown rapidly over the world, and facilitated information creation and sharing among people. Nowadays, individuals as well as organizations including governments and media outlets are leveraging social media to announce and collect local information in real time. Consequently, the immediacy and locality of information dissemination are considered important. These characteristics make social media an outstanding resource for research on text mining. By analyzing social media texts, we can understand social activities in the real world, such as trends, events, and emerging news. As a result, social media analysis has become one of the most active areas in text mining research.

Recently, the multilinguality is also considered an important characteristic in social media. However, existing works did not consider the multilinguality but instead focused on a certain language as the target of analysis. To realize cross-lingual social media analysis, it is required to solve two challenges: information sparsity and language barrier. Information sparsity is an inherent challenge due to the limited length of social media texts. To analyze social media texts, their semantic information must be expanded. In addition, social media texts are written in different languages and contain completely different terms that express the same meaning. To overcome the language barrier of social media texts, their language spaces of semantic information must be unified into a common space. In cross-lingual social media analysis, it is required to solve both challenges simultaneously.

In this thesis, we focus on cross-lingual social media analysis and solve the above challenges by using Wikipedia, a multilingual knowledge base covering comprehensive entities in the world and translatable relationships across languages. First, we introduce the research background and issues for cross-lingual social media analysis in Chapter 1.

In Chapter 2, we propose semantic relatedness measurements for social media texts.
Organizing social media texts into topics, such as text clustering, is an important task on social media analysis, and it requires to measure the semantic relatedness between texts. To solve the information sparsity and language barrier, the proposed methods generate a vector of Wikipedia related entities to an input text and compute the relatedness between social media texts by measuring the cosine similarity between generated vectors. In addition, we propose an approach to reduce the dimension of vectors to improve the time efficiency and effectiveness of the proposed methods.

In Chapter 3, we propose an entity linking method for social media texts. While social media texts are likely to be generated massively, existing entity linking methods tend to sacrifice the time efficiency to improve the effectiveness. To improve the effectiveness, the existing methods consider many entities as candidates of the corresponding entity for each mention and employ powerful but slow algorithms to compute the confidence scores for each candidate. The proposed method solves this problem by selecting a small number of meaningful candidate entities for each mention using Wikipedia's pageview.

In Chapter 4, we propose a cross-lingual trend extraction method for discovering the similarities or dissimilarities of trend information between different languages on social media. The proposed method extracts trend information as Wikipedia entities by using the entity linking method proposed in Chapter 3 to make trend information comparable between different languages. In addition, to discover the similarities, the proposed method represents entities as continuous vectors based on cooccurrences of entities in the entity linking, and finds out the similarities of trends in different languages over time by comparing the vectors.

In Chapter 5, we summarize the contributions of this thesis and discuss our future work.

様式 7

論文審査の結果の要旨及び担当者

氏　名	（　　中村　達哉　　）	
	（職）	氏　名
論文審査担当者	主査　教授	原　隆浩
	副査　教授	松下　康之
	副査　教授	藤原　融
	副査　教授	鬼塚　真
	副査　教授	下條　真司
	副査　准教授	荒瀬　由紀

論文審査の結果の要旨

Twitterなどに代表されるソーシャルメディアでは、世界中の人が自身の関心や身の回りの出来事に関する情報を自身の言語で記述した短いテキストとして多数投稿している。近年、このようなソーシャルメディアのテキストを解析し、実世界において注目されている出来事や話題を獲得するソーシャルメディア解析が注目されている。複数の言語で記述されたソーシャルメディアのテキストを解析することで、異なる言語間で共通して注目されている出来事や、ある出来事について特定の言語においてのみ注目されている観点など、言語の壁を超えた実世界の理解が可能となる。しかし、ソーシャルメディアのテキストはその短さから自身が持つ意味情報が限られており、また、異なる言語で記述されたテキストは直接比較できない。言語横断的な解析を実現するには、ソーシャルメディアのテキストが持つ意味情報を増やし、かつ、その意味情報を異なる言語間で比較可能な情報により表現する必要がある。本論文ではこれらの課題を考慮し、多言語Web百科事典Wikipediaを用いたソーシャルメディアの言語横断的な解析技術を提案している。本論文の主要な研究成果を要約すると次の通りである。

(1) 多言語かつ短文なソーシャルメディアのテキスト間の関連度計算手法を提案している。この手法では、Wikipediaの情報を用いて、任意の言語で記述されたテキストに対して関連するWikipediaの記事ベクトルを付与することで、ソーシャルメディアのテキストが持つ意味情報の拡張と異なる言語間の比較を実現している。テキストに付与されたベクトル同士を比較することにより、異なる言語で記述されたソーシャルメディアのテキスト間の関連度を計算できる。

(2) ソーシャルメディアのテキスト中の語句に対応するエンティティ（Wikipediaの記事）を紐付けるエンティティリンキング手法を提案している。既存手法では一つの語句に対して多数の記事を紐付け先の候補として考慮することで候補の取りこぼし等によるエンティティリンキングの誤りを抑制しているが、処理に要する時間が増加してしまう。そこで、紐付け先のWikipediaの記事のページビュー数を用いることにより、語句に対して候補として適切な記事を少数だけ選択し、既存手法の精度を維持したまま処理時間を高速化する手法を提案している。

(3) ソーシャルメディアからの言語横断的なトレンド抽出手法を提案している。この手法では、(2)で提案した手法を用いてソーシャルメディアのテキストと対応するエンティティを紐付けた後、エンティティが紐付けられた回数や同一のテキストに対して紐付けられたエンティティの関連性を考慮することにより、異なる言語間で同時に注目を集めているエンティティや、注目エンティティに関する観点の差異に関する分析を実現している。

以上のように、本論文はソーシャルメディアの言語横断的な解析に関する先駆的な研究として、情報科学に寄与するところが大きい。よって本論文は博士（情報科学）の学位論文として価値のあるものと認める。