氏　　名	Siriwon Taewijit
学位の種類	博士(知識科学)
学位記番号	博知第 214 号
学位授与年月日	平成 29 年 12 月 22 日
論文題目	Distant Supervision and Transductive Inference from Unlabeled Data for Medical Text Mining
論文審査委員	主査　池田　満　北陸先端科学技術大学院大学　教授
	小坂　満隆　同　教授
	Van-Nam HUYNH　同　准教授
	Dam Hieu Chi　同　准教授
	Thanaruk Theeramunkong　タマサート大学　教授

論文の内容の要旨

Text mining has been increasingly significant due to the exponential growth of data. Dealing with text mining, the text preprocessing is the additional task to transform unstructured textual data into structured one. This machine-readable format benefits not only for data comprehension, interpretation, visualization but also further utilization by traditional data mining process. Notably, relation extraction is one of the keys for discovering hidden knowledge underneath the large-scale text. The relation extraction can be though as the classification problem of a predefined relation from a given associative couple entities, for instance, the entity pair Amoxicillin-diarrhea and its relation adverse reaction. There are remaining key challenges and require manipulation by an efficient method. The supervised learning model is a well-known solution to learn attributes of a pair of entities and assigns a relation. However, the model accuracy is limited by the number of training examples, and the hand-labeled data acquisition from a large volume of text is also impracticable. While the extracting relation by pattern-based method is efficient, the manual processes for pattern generation and pattern selection are the restrictions. Moreover, the intractable processing of noisy, ill-form, domain-specific textual data and uncontrollable of unlabeled one are very challenging as well.

This dissertation mainly aims to cope with incomplete textual data (missing label) and improve the performance of relation extraction method. Regarding big data era, knowledge bases are reliable, freely available, inexpensive and maintained in multiple domains, e.g., Wikipedia for person-organization relation, SIDER for drug-event relation, IntAct for protein-protein interaction relation. The leveraging an existing knowledge base instead of manual label tagging can be seen as the promise solution for training data preparation or pattern generation, e.g., distantly supervised

relation extraction by Freebase and Wikipedia. Additionally, a key phrasal pattern is a simplified version of a given sentence but retains a semantic, e.g., *<drug>, was-held-due-to, <event>* is the phrasal pattern of the sentence *"On arrival here, propofol$_{drug}$ was held due to hypotension$_{event}$"*. The word independence assumption is widely used in Naïve Bayes for text classification due to simple but effective, although, the current word is conditionally dependent on the previous word as shown in natural language. The key phrasal pattern can benefit to reduce model complexity in the dependency representation with three elements (a drug, a pattern, an event) for all sentences instead of length l of a given sentence. Using the appropriate assumption with such data representation can yield improvement in the classification model.

To this end, the dissertation presents a framework for relation extraction from unstructured text, and the medical text will be used as a case study to extract drug-event relation. Furthermore, the dissertation introduces parameters estimation in a generative model that argues word independence assumption. This contribution can dramatically improve a model performance. Lastly, the dissertation contributes the examination on multiple approaches of incomplete data incorporation for handling unlabeled data with the efficient way.

Keywords: adverse drug reaction (ADR), medical text mining, distant supervision, multiple-instance learning (MIL), relation extraction, transductive inference

論文審査の結果の要旨

Siriwon Taewijit 君の論文は、非構造化医学文書のテキストマイニングのための遠距離教師あり学習とトランスダクティブ推定手法に関する研究をまとめたものである。

大規模文書からのテキストマイニングにおいては、文書から、着目すべき概念間の関係を抽出することが、意義のある知識を見いだす上で重要な研究課題である。医学文書においては、例えば、薬と症状の関係の記述に着目すれば、新しい薬効や副作用を見いだすことが可能になる。本論文は、着目すべき関係を表す語彙を医学文書から自動的に抽出する手法の確立を目指したものである。提案手法の特長は、前処理段階で膨大な文書からキーフレーズパターン（主語,関係語,補語・目的語）を自然言語解析と統計処理により抽出したうえで、単語間の依存関係をとらえるためのトランスダクティブ推定を適用し、パラメータ推定を行うことで、計算複雑さを抑制しつつ、従来法と比較して良好な学習結果を生成することを可能とした点にある。

具体的な応用としては、ラベルなしの非構造化テキストデータからの薬物有害反応（Adverse Drug Reaction）を検出する問題をとりあげ、遠距離教師なし学習と生成モデルにより ADR の仮説生成を実施している。結果として得られて仮説について、仮説としての医学的な意義を専門医に調査した結果、概ね ADR の可能性が認められるという評価が得られている。

さらに、本論文では、ラベルなしデータからの医学テキストマイニング手法としてよく知られている既存の方法とベンチマークデータを用いた比較実験の結果が示されている。提案手法はF1 スコアにおいて、Naïve Bayes より 11.3%, MILR 法(Multiple-instancce learning with logistic regression)より 9.3%, TSVM(Transductive support vector machine)より 9.3%の性能改善が可能であることが示されており、、提案手法が、従来法より効率的かつ有用性の高い文書データからの関係概念の抽出手法であることが示されたと言える。

本論文で示された成果は、膨大な薬品利用に関するテキストデータから、副作用・新しい処方の仮説発見に有用であり、仮説の検証を経てヘルスケア・医療分野の知識発見を導くものとして、現場応用が期待でき、知識創造プロセスの仮説生成段階の支援技術の一つとして、知識科学の新しい知見を提示していると言える。

以上，本論文は，ラベルなしデータからの医学テキストマイニングのための遠距離教師あり学習とトランスダクティブ推定による新しい関係抽出手法を提示し、その効率性・有用性を実験により実証した点において新規な知見を示しており、学術的な意義が認められる，よって博士（知識科学）の学位論文として十分価値あるものと認めた。