Title	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
Author(s)	Taewijit, Siriwon
Citation	
Issue Date	2017-12
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/15072
Rights	
Description	Supervisor:池田　満，知識科学研究科，博士

Abstract

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