氏　　名	Natthawut Kertkeidkachorn
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1971 号
学位授与の日付	平成29年9月28日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Knowledge Graph Population from Natural Language Text
論 文 審 査 委 員	主　査　准教授　市瀬　龍太郎 教授　佐藤　健 教授　武田　英明 准教授　宮尾　祐介 教授　相澤　彰子　国立情報学研究所

(別紙様式 2)
(Separate Form 2)

論文の要旨

Summary (Abstract) of doctoral thesis contents

Knowledge Graph (KG), which is a knowledge base storing knowledge in the form of real-world entities and their relationships, plays a crucial role in many modern applications, e.g. question answering, browsing knowledge, structured search, and data visualization, as prior knowledge. In recent years, there are many existing KGs such as DBpedia, Freebase, YAGO and etc. Nevertheless, it is well-known that a KG is incomplete. In addition, new knowledge regularly emerges every day. Consequently, the current KG gradually becomes more and more incomplete over time. It is therefore necessary to discover and populate new knowledge to the KG in order to fill such missing knowledge. Considering the growth of the data, we found that the volumes of unstructured data, especially natural language text, are massively exploding from various data resources. Nevertheless, natural language text has been traditionally treated as string, which cannot interpret any semantics due to the schemaless problem. Moreover, due to the complex structure of natural language text, it is not feasible for a machine to understand knowledge in natural language text. Furthermore, publishers usually publish natural language text by using their own vocabulary. It leads to the heterogeneous problem, where an identical real-world thing could be represented by many representations. Still, a large amount of knowledge in natural language text cannot be straightforwardly populated to KGs and so is left as natural language text.

Recently, there are many approaches for constructing a KG from natural language text in order to transfer knowledge to the KG. However, constructing the KG from natural language text usually builds its KG separately without integrating extracted knowledge to other existing KGs. Integrating extracted knowledge is an essential procedure because it reduces the heterogeneous problem and increases searchability over KGs. In this dissertation, we therefore aim to propose T2KG: the framework for automatically constructing/populating a KG from natural language text, where extracted knowledge is integrated to an existing KG. To integrate extracted knowledge to the existing KG, two major tasks, 1) entity linking and 2) predicate linking, are taken into account. In the framework, two sub-frameworks, namely HMiLDs and HRSim, are also proposed for dealing with the entity linking task and the predicate linking task respectively.

Linking entities to KGs becomes a challenge problem because of the continuous growth of KGs. Due to a large number of KGs, we could not know which KG contains an identical entity. As a result, some entities could not be linked to KGs. To the best of our knowledge, linking entity to multiple KGs is not addressed yet. In the entity

linking task, we therefore proposed a Heuristic expansion framework for Mapping Instances to KG data sets (HMiLDs). The main idea of HMiLDs is to directly map entities to one particular KG and then gradually expand a search space to other KGs for discovering identical entities. Due to a large amount of entities in KGs, an expansion strategy and a heuristic function for limiting the expanding search space are designed into the framework. In experiments, HMiLDs could successfully map entities to the KGs by increasing the coverage up to 90%. Moreover, experimental results also indicated that the heuristic function of HMiLDs could efficiently limit the expansion space to a reasonable space by reducing the number of candidate pairs without affecting any performances.

Predicate linking is used to identify the predicate in a KG that exactly corresponds to an extracted predicate; this is to avoid the heterogeneity problem when populating a KG. Although there have been a few studies that considered linking predicates, most of them have relied on statistical knowledge patterns, which are not able to generate the possible patterns. In the predicate linking task, we therefore proposed a Hybrid combination of Rule-based approach and Similarity-based approach (HRSim). In HRSim, we also proposed a novel distributed representation of the elements in triples and show how this can be used to compute the similarity between predicates in order to find links that would not appear in statistical patterns. The experimental results show that our distributed representation-based similarity metric outperforms other traditional similarity metrics. Also, leveraging distributed representation-based similarity metric could help to discover and identify identical KG predicates for text predicates. As a result, our approach could alleviate the problem caused by the limitation of statistical knowledge patterns due to the sparsity of text and improve the discoverability for the predicate linking task.

Finally, we introduced T2KG: the framework for populating knowledge from natural text to existing KGs. In T2KG, entity linking and predicate linking are considered when populating knowledge to existing KGs. The intuition of T2KG is to extract knowledge as triples by an open information extraction system and then integrate the knowledge into the existing KGs by performing entity linking and predicate linking in order. The experimental results show that T2KG outperforms the traditional KG construction. Although the KG population is conducted in open domains, in which any prior knowledge is not given, T2KG still achieves approximately 50% of F1 score for generating triples in the KG population task. In addition, the empirical study on the knowledge population using various text sources is conducted. The experimental results indicate T2KG could succeed to discover new knowledge that does not exist in DBpedia.

(別紙様式 3)
(Separate Form 3)

博士論文審査結果の要旨

Summary of the results of the doctoral thesis screening

知識グラフは，現実世界のエンティティとその関係を記述した知識ベースであり，質疑応答や検索など様々な知的処理で利用されている．しかし，日々，新しい知識が生ずるため，新たな知識を知識グラフに追加していく手法が必要となる．博士論文では，自然言語で書かれた文書を用いて知識グラフに知識を追加する問題について取り組んでいる．知識を統合するために必要なエンティティ結合，述語結合の 2 つの課題と，実際に知識グラフに知識を追加する課題の 3 つに課題を分け，それらの課題を解決したというのがこの論文の主張である．

本論文は，全 7 章からなる．第 1 章「Introduction」では，知識グラフに知識を追加する問題に対して，研究の背景を述べると共に，本論文の貢献について説明している．

第 2 章「Fundamentals and Related Work」では，本論文の前提となる基礎的な事項と関連研究についてまとめている．最初に，知識グラフやそれを生成する手法について説明している．その後，エンティティ結合，述語結合を含めた知識グラフの統合についての関連研究を述べ，最後にこれまでの研究で残された課題について議論し，本研究の取り組む課題の明確化を行っている．

第 3 章「Entity Linking」では，エンティティ結合の問題とその解決法について述べている．複数の知識グラフの中から，同じエンティティを探す問題を提起し，候補選択コンポーネントなどで構成される解決のためのフレームワーク HMiLDs を提案している．そして，4 つの実験を通して検証を行い，提案手法の有効性を示している．

第 4 章「Predicate Linking」では，述語結合の問題とその解決法について述べている．知識を追加する際に，知識グラフに同じ意味の述語が複数存在することを避けるために，同じ意味の述語を同定する問題を提起し，2 つの手法を組み合わせて解決するフレームワーク HRSim を提案している．そして，実験的に検証を行い，提案手法の有効性を示している．さらに，この問題にも応用できるベクトルを用いた語の表現方法を提案し，その実験的な有効性も示している．

第 5 章「Knowledge Graph Population」では，自然言語で書かれた文書から知識グラフに知識を追加する問題とその解決法について述べている．自然言語で書かれた文書から知識グラフに知識を追加するための，フレームワーク T2KG を提案している．このフレームワークは，エンティティ結合，述語結合を行うコンポーネントを含めて構成されている．そして，3 つの実験を通して検証を行い，提案手法の有効性を示している．

第 6 章「Discussion」では，提案した 3 つのフレームワーク HMiLDs, HRSim, T2KG について総合的な考察を行い，この論文の到達点，限界点などを議論している．

第 7 章「Conclusion」では，博士論文の総括を行うと共に，展望を述べ，結論をまとめている．

上記のように，本博士論文は，自然言語で書かれた文書から知識グラフに自動的に知識を追加するための 3 つの課題を提示し，その解決手法を示した点で，この研究分野の発展に貢献するものである．また，本論文で示した考え方により，新たな知識を自動的に既存の知識グラフに追加できることが示されており，知識処理のための基盤技術開発という観

(別紙様式 3)
(Separate Form 3)

点からも意義があると認められる．さらに，博士論文の内容は，1 本の査読付きジャーナル論文，3 本の査読付き国際会議論文，1 本の査読付き国際ワークショップ論文などで発表されており，社会からも評価されている．以上より，本論文は博士論文として，十分な水準であると審査委員全員一致で認められた．