氏　　名　Esrat Farjana Rupu

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2454 号

学位授与の日付　2023 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 Extraction for Open Knowledge Graph under Open World Assumption

論 文 審 査 委 員　主　査　武田 英明
情報学コース 教授
佐藤 健
情報学コース 教授
相澤 彰子
情報学コース 教授
菅原 朔
情報学コース 助教
市瀬 龍太郎
東京工業大学 工学院 教授

Summary of Doctoral Thesis

Name in full: Esrat Farjana Rupu

Title: Knowledge Extraction for Open Knowledge Graph under Open World Assumption

The usefulness and usability of existing knowledge graphs (KGs) are mostly limited because of the incompleteness of knowledge compared to the growing number of facts about the real world. Most existing ontology-based KG completion methods are based on the closed-world assumption, where KGs are fixed. In these methods, entities and relations are defined, and new entity information cannot be easily added. In contrast, in open-world assumptions, entities and relations are not previously defined. Thus, there is a vast scope to find new entity information. Despite this, knowledge acquisition under the open-world assumption is challenging because most available knowledge is in a noisy unstructured text format. It is not possible to extract entity information directly from the natural text because it is unstructured. In this context, the Open Information Extraction (OpenIE) system extracts binary relationships in the triple format (e.g., (Barack Obama, was born in, Honolulu)) from unstructured text without any prespecified vocabulary. Although OpenIE does not require any prior knowledge, the quality of OpenIE triples varies. The system is likely to include lots of noisy and redundant information in KBs, making them inconsistent. The thesis concerns the method to extract the necessary information for KGs from a triple set those are created by an OpenIE tool. As extracted triples by openIE are noisy, therefore, to use such triples for the KG, it is necessary to build a system that can identify triples for KGs from the extracted noisy triple set.

To be specific, the thesis has proposed a method for identifying competent triples from a given triple set by the OpenIE tool. Here, competent triples are the triples that can use to build an open KG as well as can contribute to adding new information to the existing KGs. As far as we know, this is the first work to identify competent triples that are extracted from natural text using the OpenIE tool.

In this thesis, we propose the Competent Triple Identification (CTID) model for KGs. We also propose two types of features, namely syntax- and semantic-based features, to identify competent triples from a triple set extracted by a state-of-the-art OpenIE system. We investigate both types of features and test their effectiveness. It is found that the

performance of the proposed features is about 20% better compared to that of the ReVerb system in identifying competent triples.

In this thesis, chapter 1 contains the background and a brief introduction of the knowledge graph with examples. The structure and use cases of KGs are briefly explained in this chapter. Here, we can also learn about some well-known KGs and what is their role in many advanced applications. Chapter 1 also briefly introduces the motivation and contributions of the proposed method in this thesis.

Chapter 2 presents the fundamentals of the knowledge graph and its concept and knowledge representation. Then, the background knowledge for knowledge graph construction. Later, we further discussed the natural language models that frequently uses in the KG completion tasks that are focused on text data. We also discuss the OpenIE system to extract triples from natural text. Finally, we discussed closed and open-world assumptions before concluding the chapter.

Chapter 3 presents the details of the proposed CTID model for competent triple identification from natural text using the OpenIE tool. Here, our proposed features are discussed elaborately. In the CTID model, we proposed two types of feature sets namely syntax- and semantic-based features. In this chapter, we briefly explained these two types of feature sets and how these features contribute to extracting triples for the KGs. Next, we include the experiment conducted to evaluate the proposed features. Here, we utilize a QA dataset to create our dataset for the evaluation. We also include an ablation study and some limitations here. Finally, we discuss and summarize the CTID model.

Chapter 4 discusses the limitations and the achievement we accomplished by the proposed model using two different types of feature sets. Here, we include additional experiments which include human label annotation and compare the result which is built by using our annotation algorithm. We can understand the limitations of our annotation algorithm from this further analysis. In this chapter, we also discuss some additional parameters that are used in our proposed model and how to choose those parameters and what is the effect of those parameters. Here, we also include the assumptions of our research briefly.

Chapter 5 concludes the thesis and summarizes the thesis's contributions and outlines future work directions.

（別紙１ Separate Sheet 1）

Results of the doctoral thesis defense

博士論文審査結果

Name in Full
氏　　名　　Esrat Farjana Rupu

Title
論文題目　　Knowledge Extraction for Open Knowledge Graph under Open World Assumption

知識グラフは，有用な知識ベースとして質疑応答や推薦システムなど，様々なタスクで利用されている．そのため，知識グラフに，不足している知識を補完する手法が多く研究されている．既存の知識グラフ補完の手法の多くは，知識グラフで使われるエンティティや関係が予め定義され，新しいエンティティなどの情報を追加することが困難な問題があるが，開世界仮説（Open World Assumption）の下では，エンティティや関係が事前に定義されていないため，多くの新たな知識を補完することが可能となる．本論文では，開世界仮説の下で，知識グラフに適した知識をテキストから抽出する手法を確立することを目的とする．そのために，テキストから OpenIE を利用して情報抽出を行った後に，知識グラフに適切な知識を識別するという課題について取り組んだ．

本学位論文は，全５章からなる．第１章「Introduction」では，本研究で取り組んだ問題に対する動機，背景について述べると共に，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，知識グラフに関連する基礎的な概念，本研究で用いた OpenIE システムなど，本論文を理解するために必要となる事項について説明すると共に，本論文で重要な概念となる開世界仮説について説明を行っている．

第３章「Competent Triple Identification」では，OpenIE を利用して自然言語で書かれたテキストから，知識グラフに適切な知識を特定するための手法，CTID モデルを提案し，その手法の有効性を実験的に示している．

第４章「Discussion」では，本論文で取り組んだ課題について，提案手法の問題点を分析し，本研究の到達点を明らかにしている．

最後の第５章「Conclusion」では，博士論文の総括を行うと共に，結論をまとめている．

公開発表会では，博士論文の章立てに従って発表が行われた．その後に行われた論文審査会，及び，口述試験では，審査委員からの質疑に対して的確に回答がなされた．

質疑応答の後に，審査委員会を開催し，審査委員で審議を行った．審査委員会では，出願者の博士論文で，知識グラフに適した知識を開世界仮説の下でテキストから抽出する手法を構築した点で評価された．また，提案手法により，知識グラフの知識補完方法の適用範囲が，従来よりも広がるため，基盤技術開発という観点からも評価された．

以上を要するに，本学位論文は，知識グラフに関して，新たな知識補完方法を示したものであり，研究分野の発展に貢献しているという点で学術的価値が大きい．また，本学位論文の成果は，学術雑誌論文１件，国際会議論文１件として発表され，社会的な評価も得ている．以上の理由により，審査委員会は，全員一致で本学位論文が学位の授与に値すると判断した．