| | |
|---|---|
| 氏　　名 | Rumana Ferdous Munne |
| 学位(専攻分野) | 博士(情報学) |
| 学 位 記 番 号 | 総研大甲第 2412 号 |
| 学位授与の日付 | 2023 年 3 月 24 日 |
| 学位授与の要件 | 複合科学研究科　情報学専攻<br>学位規則第6条第1項該当 |
| 学 位 論 文 題 目 | Entity Alignment and Attribute Enhancement between Knowledge Graphs |
| 論 文 審 査 委 員 | 主　査　山田 誠二<br>情報学専攻　教授<br>佐藤 健<br>情報学専攻　教授<br>相澤 彰子<br>情報学専攻　教授<br>武田 英明<br>情報学専攻　教授<br>市瀬 龍太郎<br>東京工業大学　工学院　教授 |

(Form 3)

# Summary of Doctoral Thesis

Name in full Rumana Ferdous Munne

Title Entity Alignment and Attribute Enhancement between Knowledge Graphs

A Knowledge Graph (KG) is a knowledge model containing facts about real world entities represented as a graph. It is a collection of interlinked descriptions of entities, relationships, concepts, and events. We have witnessed rapid growth in knowledge graph creation and application in last few years. Several efforts have been made to develop knowledge graphs in
general and specific domains such as DBpedia, YAGO, LinkedGeoData, and Wikidata and they have been served several fields of real-world applications from semantic parsing and named entity disambiguation to information extraction and question answering. These knowledge graphs contain millions of facts about entities. However, these knowledge graphs are far from complete and mandate continuous enrichment and enhancement. One possible approach to enhance KG is integrating knowledge from various knowledge graphs based on their aligned information. In this thesis, we develop new effective methods to find aligned entities from different KGs first and later enrich the KGs by enhancing their attributes.

We start this thesis by presenting techniques for entity alignment. The task of entity alignment is to find entities in two heterogeneous knowledge graphs that represent the same real-world entity. Many knowledge graphs have been created separately for certain purposes with overlapping entity coverage. These knowledge graphs are complementary to each other in terms of completeness. Unfortunately, only a fraction of the entities stored in different KGs are aligned.

We present an embedding-based entity alignment method that finds entity alignment by estimating the similarities between entity embeddings. Existing methods mainly focus on the relational structures and attributes information for the alignment process. Such methods fail while the entities have limited number of attributes or when the relational structure couldn't capture the meaningful representation of the entities. To overcome this problem, we propose EASAE, an Entity Alignment method using Summary and Attribute Embeddings. We utilize the entity summary information available in KGs for entities' summary embedding by employing BERT. Our model learns the representations of entities by using relational triples, attribute triples, and summary as well. Extensive experiments show that entity summary exhibit useful semantic information in entity alignment task. Our proposed approach outperforms the concurrent state-of-the-art alignment models.

To enrich KG by enhancing their attributes, we propose an Attribute Enhancement

Framework (AEF) that integrates multiple KGs based on their aligned information. Typically, similar entities from different KG contain different set of attributes. AEF exploits representation learning based ranking model to discover the significant attributes from reference KG. Later it employs a similarity mapping technique to integrate new attributes into the target KG. AEF also determines the attribute value inconsistency between two KGs. With this study, we aim to include all the essential attributes to the existing KGs towards more robust and complete knowledge graph. The results of the attribute enhancement work indicate important directions for the future work.

Knowledge is the core power in the age of data and information and incorporating the knowledge in a graphical representation is known as Knowledge Graph (KG). It facilitates the complex process of searching and exploration as a lot of information is in the form of data, and context about an entity, or object. We intend to enrich and enhance KG which in turn helps to represent human knowledge into structured models that plays a vital role in the machine learning domain.

The knowledge graph is a structured representation of real-world knowledge, consisting of entities, relationships, attributes, and textual descriptions. An entity or instance is an object in the real world; a relationship describes the interaction and relation between two entities; an attribute describes the property of an entity; and a textual description includes the entity abstract, short summary, string information, etc. Knowledge graph is modeled by a graph structure and facts are mainly represented in triple format. A triple consists of a subject, a predicate/relation, and an object where the predicate/relation indicates the relationship between an entity as the subject and the other entity (or literal) as the object. A relational triple can be denoted as *(h, r, t)* where *h* and *t* are the head entity and the tail entity respectively and *r* is the relation between the *h* and *t*. If the object is literal, then we call it attribute triple.

There has been a surge of research and development on KGs for several decades due to their effective role in storing and representing knowledge and facts. Several well-known KGs can be found on the web including open-source ones such as DBpedia, Freebase, YAGO, as well as commercial ones such as those developed by Google and Microsoft. In the last few years, there has been an explosive growth of interest in KGs in both the research community and the industry due to their indispensable role in AI applications such as natural language processing (including dialogue systems/chatbots, question answering, sentence generation, etc.), search engines, recommendation systems, and information extraction.

KGs are mostly constructed based on crowd-sourced content and automatic extraction methods; therefore, they are not always complete and error-free. Also, different KGs are initially constructed independently to serve various purposes, and they are heterogeneous by their structure. Integrating multiple KGs can complement each other toward a more complete and robust knowledge representation which requires KG

alignment process. As a result, KG alignment and knowledge enhancement task become very prominent research area.

Over the last decade, several knowledge graphs have been created though most of them have significant overlapping entity coverage. These knowledge graphs e.g. DBpedia, YAGO, Wikidata are complementary to each other in terms of completeness. This study aims to integrate facts belong to different knowledge graphs to form more robust knowledge graphs. Before we can incorporate multiple KGs, it is required to align them first. Aligning multiple KGs means we have to align the overlapping entities. Entity alignment task refers to finding the same real-world entities in multiple KGs. The existing entity alignment models mostly depend on relational and attribute triples. This is because the neighbors of two equivalent entities in KGs usually contain equivalent entities and two equivalent entities often share similar attributes and values in KGs. However, these neighboring assumptions become inadequate when we have to deal with 1-N or N-N relations. Also, entity pairs from two KG will not always share equivalent attributes. It is very common for different KGs entities to contain a different set of attributes.

Recently embedding-based approaches are popular for entity alignment task. Such models are built on top of a graph embedding model, which learns entity embeddings that capture the similarity between entities in a knowledge graph based on the relationship triples in a KG. To adapt the KG embedding for entity alignment between two KGs, the embedding-based models require both predicate and entity embeddings of two KGs to fall in the same vector space. Exiting models mostly rely on large numbers of seed alignments for this. However, the seed alignments between two KGs are rarely available, and hence are difficult to obtain due to the expensive human efforts required. We have exploited predicate alignment to ease the situation. We propose a novel model for embedding-based entity alignment which utilizes entity summary extracted from the KGs along with the relational and the attribute triples. Summary, relation, and attribute embeddings are jointly optimized to improve the alignment performance.

The goal of this research work is to integrate multiple KG towards a more complete and robust KG. Technically, when the KGs are aligned, they can complement each other both in terms of relations and attributes. However, traditional KG completion techniques are quite popular for solving relation completeness problem. Several probabilistic, embedding and deep learning models are generally used for the completion of relations. Completion of attributes on the other hand is a more complex task. It is often challenging to detect the existence of missing attributes. Even if we can discover the missing attributes, the classification method may be challenging to apply widely due to the different hierarchical structures of different entities. Exiting models on attribute completion focuses on extracting attributes from the text description or source associated with specific entities mostly. This process works well

when missing or incomplete attributes occurs due to inadequate data extraction method. But the issue still remains if the data source is not complete or the overall structure of KG lacks many important properties. In such cases, incorporating multiple KGs for attribute completion can be very useful. So we have proposed an attribute-enhancing framework that attempts to enrich the attributes of entities by integrating multiple KGs based on their aligned information. The task is done by first detecting significant attributes from reference KG and exporting them to corresponding entities from target KG if target KG is missing that information.

In this dissertation, we aim to incorporate multiple heterogeneous knowledge graphs which in terms lead us toward more complete KGs. We address the problem of entity alignment which is a vital process for KG integration and propose some techniques to tackle that task. We also discuss the knowledge integration processes and how we can solve them with a simple solution.

Form8 : Separate Sheet
（様式８・別紙１）

**Results of the doctoral thesis screening**

# 博士論文審査結果

Name in Full
氏　　名　　Rumana Ferdous Munne

Title
論文題目　　Entity Alignment and Attribute Enhancement between Knowledge Graphs

知識グラフ（Knowledge Graph）は，実世界のエンティティに関するファクトをグラフで表現した知識モデルである．既存の知識グラフには，多くの知識が含まれるが，継続的な知識の増強が必要となる問題がある．その問題を解決するため，本論文では，様々な知識グラフから得られるファクトを統合することで，知識を増強する手法を確立することを目的とする．そのために，２つの課題に取り組んだ．まず，（１）異なる知識グラフから同じものを指すエンティティを発見(Entity Alignment)し、その後、（２）属性情報を他の知識グラフから取り入れることで知識グラフを増強(Attribute Enhancement)させる．

本学位論文は，全６章からなる．第１章「Introduction」では，本研究で取り組む課題である知識グラフの増強問題に関して，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 Works」では，本論文の理解に必要となる知識グラフ埋め込みなどの基礎事項を説明すると共に，関連研究について述べている．

第３章「Entity Alignment via Summary and Attribute Embeddings」では，１番目の課題となる Entity Alignment に関して，知識グラフに含まれるエンティティの要約情報を利用する新たな手法を提案し，その有効性を実験的に示している．

第４章「Attribute Enhancement using Aligned Entities between Knowledge Graphs」では，２番目の課題となる Attribute Enhancement に関して，重要な属性を発見して統合を行う新たな手法を提案し，その有効性を実験的に示している．

第５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．

第６章「Conclusion」では，博士論文の総括を行うと共に展望を述べ，結論をまとめた．

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，審査委員で議論を行った．審査委員会では，出願者の博士研究が，２つの課題において，新たな視点を用いた有効な手法を示した点で評価された．また，本論文では，知識グラフ構築の際に，大きな課題となっていた知識の増強方法に関して，新たな展望を示しており，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，知識グラフの増強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，本学論文の成果は，学術雑誌論文１件，査読付き国際会議論文２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．