KG Quality Metrics

Student Name: Pranav Sharma
Roll Number: 2020395

BTP report submitted in partial fulfillment of the requirements
for the Degree of B.Tech. in Computer Science & Engineering
on ...(27 November 2023)...

BTP Track: Research Track

BTP Advisor
Dr Raghava Mutharaju
Dr Manuj Mukherjee

Indraprastha Institute of Information Technology
New Delhi

Student's Declaration

I hereby declare that the work presented in the report entitled **"KG Quality Metrics"** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Engineering* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Raghava Mutharaju** . Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

..... **Place & Date: IIITD, New Delhi (27 May 2023)**
Pranav Sharma

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

..... **Place & Date: IIITD, New Delhi (27 May 2023)**
Dr. Raghava Mutharaju

Abstract

In the burgeoning field of knowledge representation, the construction and maintenance of high-quality knowledge graphs (KG's) play a pivotal role in ensuring the accuracy and reliability of information. This research endeavors to establish a comprehensive framework for assessing the quality of knowledge graphs, introducing novel matrices and metrics tailored to capture the intricacies of knowledge representation. Our approach involves the development of quantifiable measures that evaluate aspects such as completeness, consistency, accuracy, and contextual relevance within a knowledge graph.

Keywords: Knowledge graph (KG), Quality assessment, Novel matrices, Accuracy, Completeness, Consistency, Contextual relevance

Acknowledgments

I would like to express my sincere gratitude to my Professor and our project supervisor, Dr. Raghava Mutharaju and Dr. Manuj Mukherjee, for their invaluable guidance and support throughout this project. Working with them has been a fantastic experience, as they have mentored me in the field of Knowledge Graphs and Evaluation Matrices, and have provided insights and expertise at every stage of the project. Their commitment and dedication have been instrumental in achieving our project goals, and I look forward to continuing our collaboration in the future. I would also like to thank the team members who provided support and encouragement, making this project an enriching experience

Contents

Chapter 1

Introduction

In today's era of big data, the need for efficient and accurate information retrieval has become more critical than ever. To address this challenge, knowledge graphs have emerged as a powerful tool for organizing and structuring complex data. Knowledge graphs are composed of nodes and edges that represent entities and the relationships between them. By analyzing these relationships, knowledge graphs enable users to gain insights into complex data sets quickly and easily.

Open KGs, such as DBpedia, Freebase, and Wikidata, democratize access to knowledge, making it freely available for public exploration and utilization. These open-source repositories foster collaboration and innovation, empowering researchers, developers, and enthusiasts to expand the boundaries of knowledge representation. [3]

Conversely, enterprise KG's reside within the confines of organizations, catering to specific commercial needs. These internal knowledge bases drive business intelligence, enabling companies to gain insights from their data, enhance customer experiences, and optimize operations. The likes of Bing, Google, Airbnb, and Facebook harness the power of enterprise KGs to refine their search algorithms, personalize recommendations, and connect users to relevant information and services.

Whether open or enterprise, KGs play a pivotal role in transforming raw data into structured knowledge, paving the way for intelligent systems that can navigate the complexities of the real world.[3]

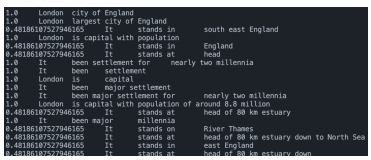

Figure 1.1: StanfordIE

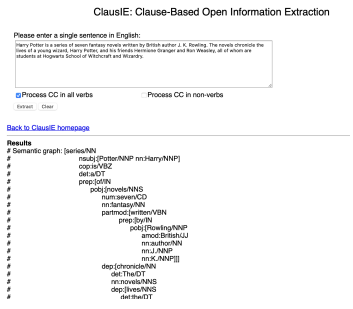

Figure 1.2: CluasIE

Chapter 2

Literature Review

In the realm of data management, the shift towards graph data models has gained significant traction due to their inherent flexibility and adaptability in handling diverse data sources. Relational database schemas, while prevalent in traditional data management, often face limitations in representing complex relationships and evolving data structures. Graph data models, on the other hand, excel in capturing intricate connections between entities and accommodating dynamic data changes.

Relational Database Schema Limitations

Relational database schemas, the cornerstone of traditional data management, rely on a structured, tabular format to store and organize data. This approach, while effective for certain types of data, often falls short when dealing with complex relationships and evolving data needs. For instance, modeling intricate connections between entities, such as relationships between people, organizations, and events, can become cumbersome in a relational database schema. Moreover, the rigid structure of relational databases can hinder the integration of new data sources, requiring upfront schema modifications that can be time-consuming and error-prone.

Advantages of Graph Data Models

Graph data models offer a compelling alternative to traditional relational models, providing a more flexible and adaptable framework for representing and managing complex data. Unlike relational databases, graph data models utilize a network-based approach, where entities are represented as nodes and relationships between entities are represented as edges. This approach offers several advantages, including:

Flexibility: Graph data models can accommodate a wide range of data structures, including hierarchical, cyclic, and n-ary relationships. This flexibility makes them well-suited for modeling complex real-world relationships.

Adaptability: Graph data models can evolve seamlessly to incorporate new data sources and adapt to changing data requirements. This adaptability is particularly valuable in dynamic environments where data is constantly evolving.

Expressiveness: Graph data models can capture intricate relationships and context, enabling more meaningful and nuanced data analysis. This expressiveness is crucial for understanding complex systems and making informed decisions.

Real-World Application: Tourism Board's Data Management

The tourism board's experience provides a compelling illustration of the limitations of relational database schemas and the potential of graph data models. Initially, the tourism board relied on

a relational database to manage their data, including information about events, attractions, and accommodations. However, as the volume and diversity of their data increased, the relational database became increasingly difficult to maintain and query.[3]

Stanford OpenIE is an open-source Java implementation of an Open Information Extraction (OIE) system developed by the Stanford Natural Language Processing Group. It is a tool that extracts relation triples, typically binary relations, from plain text. For example, it can extract the triple (Mark Zuckerberg; founded; Facebook) from the sentence "Mark Zuckerberg founded Facebook."[1]

ClausIE is a novel approach to open information extraction (OIE) that extracts relations and their arguments from natural language text. ClausIE differs from traditional OIE systems in that it operates at the clause level, rather than the sentence level. This allows ClausIE to capture more fine-grained information about the relationships between entities.[2]

OpenIE6 is a state-of-the-art open information extraction (OIE) system that utilizes an iterative grid labeling and coordination analysis approach. It excels in extracting comprehensive and accurate relation triples from natural language text.[4]

Chapter 3

Background Study

Knowledge graphs (KGs) have emerged as a powerful tool for representing and managing knowledge in a structured and machine-readable format. They are increasingly being used in a variety of applications, including search engines, recommendation systems, and virtual assistants. However, evaluating the quality of KGs is a challenging task, as there is no single metric that can capture all aspects of KG quality.

Several metrics have been proposed for evaluating KG quality, but they typically fall into two main categories: structural metrics and semantic metrics. Structural metrics focus on the overall structure of the KG, such as the number of entities and relationships, the average path length between entities, and the distribution of relationship types. Semantic metrics focus on the meaningfulness of the information in the KG, such as the accuracy of entity names and relationships, the completeness of the KG, and the consistency of the KG with other knowledge sources.

One way to evaluate the effectiveness of IE algorithms is to compare them to an ideal form of a KG. This ideal KG would be complete, consistent, and accurate. However, it is impractical to create a perfect KG, so researchers have instead used approximations of an ideal KG, such as manually curated KGs or KGs extracted from gold standard datasets.

By comparing IE algorithms to an ideal KG, researchers can assess the ability of the algorithms to extract high-quality information from text. This can help to identify the strengths and weaknesses of different IE algorithms and inform the development of new and improved algorithms.

In addition to comparing IE algorithms to an ideal KG, researchers can also evaluate the quality of KGs using a variety of other methods. These methods include using human experts to evaluate the quality of KGs, using automated tools to assess the quality of KGs, and using KGs in real-world applications to assess their impact.

The development of effective methods for evaluating KG quality is an important area of research. By developing better methods for evaluating KGs, researchers can help to ensure that KGs are of high quality and can be used effectively in a variety of applications.

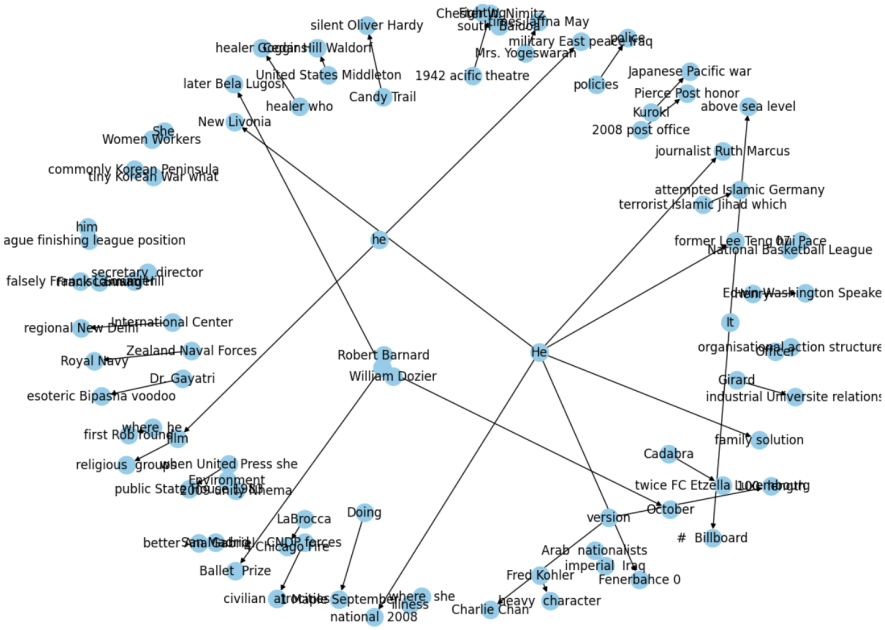

Figure 3.1: Graphs By Spacy Library

Chapter 4

Analysis

The provided information describes a comprehensive approach to evaluating the performance of three open information extraction (OIE) models. The evaluation process involves comparing the extracted triplets from each model to an ideal triplet form, which represents the perfect extraction for a given sentence. This comparison allows for the assessment of various factors, including the number of extracted triplets, the completeness of sentence coverage, and the accuracy of extracted relationships.

To further enhance the evaluation, the study considers the transformation of nouns into nodes within the extracted triplets. This transformation provides insights into how effectively each OIE model identifies and represents entities within the text. By comparing the number of noun-based nodes across different models, the study can identify potential advantages or limitations in entity recognition capabilities.

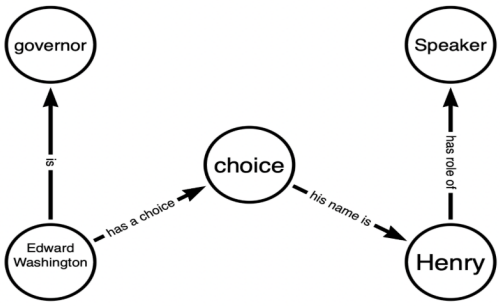

Figure 4.1: Henry was Governor Edwin Washington Edwards's choice for Speaker.

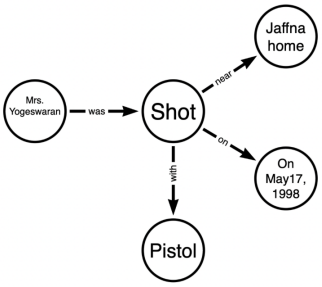

Figure 4.2: Mrs. Yogeswaran was shot five times with a pistol near her Jaffna home on May 17, 1998.

The inclusion of manually drafted ideal knowledge graph forms adds another layer of evaluation, allowing for a more holistic assessment of the OIE models. By comparing the extracted triplets to these manually curated knowledge graphs, the study can identify areas where the models deviate from human-generated knowledge representations. This comparison can reveal potential biases or limitations in the models' extraction strategies.

Overall, the described approach provides a rigorous and comprehensive framework for evaluating the performance of OIE models. By considering various factors, including triplet formation, sentence coverage, entity recognition, and comparison to ideal knowledge graphs, the study can

identify the strengths and weaknesses of each model and contribute to the development of more accurate and effective OIE systems.

Chapter 5

Conclusion

The evaluation of knowledge graphs (KGs) extracted from open information extraction (OIE) algorithms requires a comprehensive approach that considers various factors. This study proposes a novel framework that incorporates multiple metrics to assess the quality of KGs.

One key metric is the number of triplets extracted for each sentence. A higher number of triplets indicates more comprehensive coverage of the sentence's information. However, this metric alone is insufficient, as it does not account for the accuracy or completeness of the extracted relationships.

To address this limitation, the study introduces a completeness measure that evaluates whether the extracted triplets capture the entirety of the sentence's meaning. This measure ensures that the extracted information is not fragmented or incomplete.

Furthermore, the study proposes comparing the extracted KGs to manually drafted ideal knowledge graphs. This comparison allows for the assessment of the OIE models' ability to accurately represent the relationships between entities in the text.

To further refine the evaluation process, the study suggests converting the extracted triplets into numerical vectors and analyzing their proximity to ideal knowledge graph vectors. This approach provides a quantitative measure of the accuracy of the extracted relationships.

Measure of Correctness

The measure of correctness quantifies the distribution of information within the extracted knowledge graphs. This measure assesses whether the information is evenly distributed across the graph or concentrated in a few nodes. A higher correctness value indicates a more balanced distribution of information, implying that the knowledge graph effectively represents the relationships between entities in the text.

The correctness measure is calculated by averaging the following formula for each triplet graph of a sentence:

Correctness = (Nouns in each node) / Total number of nodes in the graph
This formula essentially calculates the average number of nouns per node in the graph. If this value is close to one, it indicates that the information is well-distributed across the graph. Conversely, if the value is significantly lower than one, it suggests that the information is concentrated in a few nodes, potentially leading to inaccuracies or incomplete representation of the relationships. By incorporating these multiple metrics, the proposed framework provides a comprehensive and rigorous approach to evaluating the quality of KGs extracted from OIE algorithms. This approach can help identify the strengths and weaknesses of different OIE models

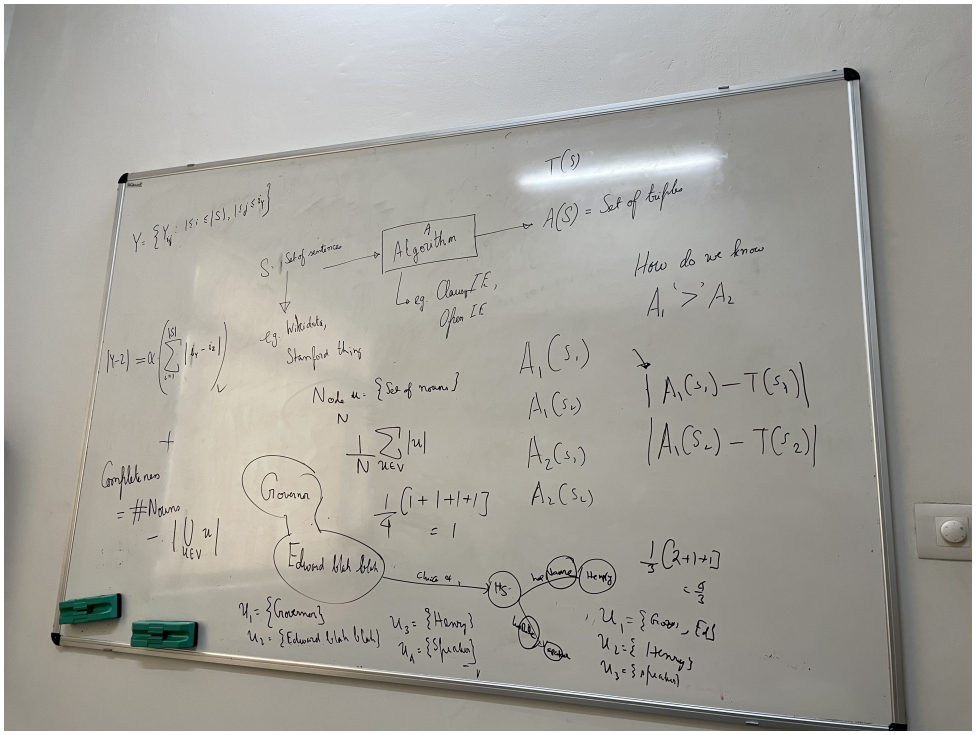

Figure 5.1: Empirical Research

and inform the development of more accurate and effective IE systems.

Bibliography

[1] Addison, A., and Gaiani, M. Virtualized architectural heritage: new tools and techniques. *IEEE MultiMedia 7*, 2 (2000), 26–31.

[2] Del Corro, L., and Gemulla, R. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd International Conference on World Wide Web* (New York, NY, USA, 2013), WWW '13, Association for Computing Machinery, p. 355–366.

[3] Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. Knowledge graphs. *ACM Computing Surveys 54*, 4 (July 2021), 1–37.

[4] Kolluru, K., Adlakha, V., Aggarwal, S., Mausam, and Chakrabarti, S. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online, Nov. 2020), B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Association for Computational Linguistics, pp. 3748–3761.