Building Knowledge Graphs from Unstructured Text

Team lead: Sohaib Zafar (p20-0574)

Research Team: Usama Ahmad (p19-0094) and Hassan Ahmad (p18-0120)

Data Collection Team: Meer Murtaza (p18-0145) and Azam Irfan (p20- 0154)

Code Team: Sohaib Zafar (p20-0574), Abu Huraira (p20-0612), Umer Saleem (P20-0136), Usama Asghar (P19-0092), Amir Hamza (P19-0059), Fahad Nawaz P(19-0037)

Documentation Team: Hafiz Usman (p18-0100) and Muhammad Ali (p17- 6065)

***Abstract*—The availability of vast amounts of unstructured text data on the web has led to the need for methods to extract structured knowledge from it. Knowledge graphs have emerged as a popular way to represent structured knowledge, and there has been significant research in building them from unstructured text. This paper presents a survey of the current state-of-the-art techniques for building knowledge graphs from unstructured text. We cover the major steps involved in this process, including data acquisition, text pre-processing, entity and relationship extraction, graph construction, and knowledge representation. We also discuss the challenges and open research problems in this area. Our survey includes a comprehensive review of recent research papers, highlighting the key contributions and limitations of each approach. Finally, we provide a comparative analysis of the surveyed techniques and discuss potential future directions for research in this field.**

***Index Terms*—structured, data acquisition, text pre-processing, entity**

I. Introduction

The availability of vast amounts of unstructured text data on the web has led to the need for methods to extract structured knowledge from it. Knowledge graphs have emerged as a popular way to represent structured knowledge, and they have been used for a variety of applications such as question answering, recommendation systems, and natural language understanding. Building knowledge graphs from unstructured text involves extracting relevant information from text and representing it in a structured format. This process involves several steps, including data acquisition, text pre-processing, entity and relationship extraction, graph construction, and knowledge representation.

In recent years, there has been significant research in building knowledge graphs from unstructured text. Researchers have proposed a wide range of techniques to address the challenges involved in this process, such as the ambiguity and variability of natural language, the diversity of data sources, and the scalability of the approach. However, despite the progress made in this area, several challenges remain, including the need for more accurate and efficient techniques, the need to handle multi-lingual and multi-modal data, and the need for better evaluation metrics.

This paper presents a survey of the current state-of-the-art techniques for building knowledge graphs from unstructured text. We provide a comprehensive review of recent research papers and highlight the key contributions and limitations of each approach. We also discuss the major steps involved in building knowledge graphs from unstructured text and the challenges and open research problems in this area. Finally, we provide a comparative analysis of the surveyed techniques and discuss potential future directions for research in this field. This survey aims to provide a comprehensive overview of the current state of research in building knowledge graphs from unstructured text and to help researchers and practitioners in this field to identify promising approaches and research directions.

Identify applicable funding agency here. If none, delete this.

II. Methodology

To build a knowledge graph from text, we typically need to perform two steps: Extract entities, a.k.a. Named Entity Recognition (NER), which are going to be the nodes of the knowledge graph. Extract relations between the entities, a.k.a. Relation Classification (RC), which are going to be the edges of the knowledge graph. These multiple-step pipelines often propagate errors or are limited to a small number of relation types. Recently, end-to-end approaches have been proposed to tackle both tasks simultaneously. This task is usually referred to as Relation Extraction (RE). In this article, we'll use an end-to-end model called REBEL, from the paper Relation Extraction By End-to-end Language generation.

A. How REBEL Works

REBEL is a text2text model trained by BabelScape by fine-tuning BART for translating a raw input sentence containing entities and implicit relations into a set of triplets that explicitly refer to those relations. It has been trained on more than 200 different relation types. The authors created a custom dataset for REBEL pre-training, using entities and relations found in Wikipedia abstracts and Wikidata, and filtering them using a RoBERTa Natural Language Inference model (similar to this model). Have a look at the paper to know more about the creation process of the dataset. The authors also published their dataset on the Hugging Face Hub. The model performs quite well on an array of Relation Extraction and Relation Classification benchmarks.

B. Steps taken

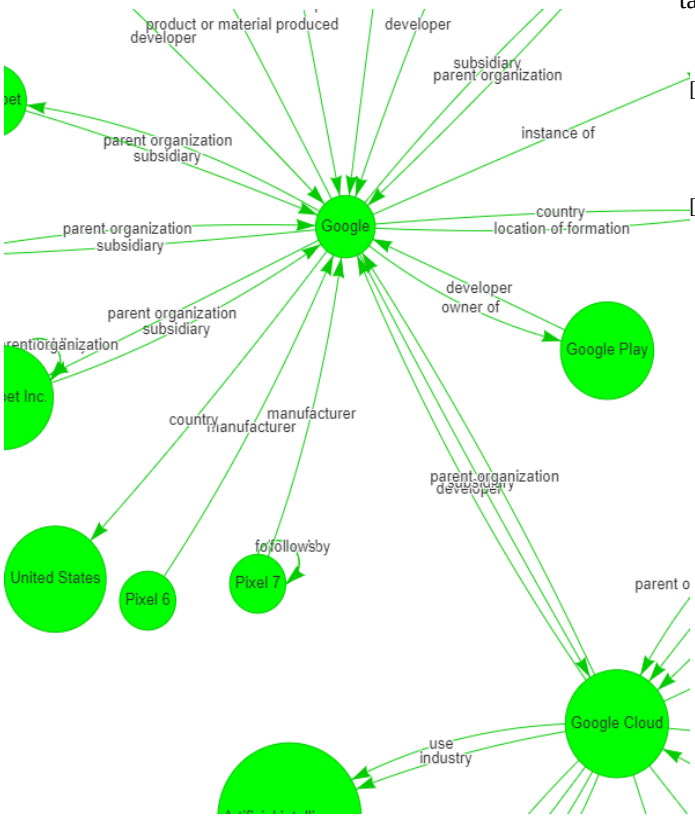
- Load the Relation Extraction REBEL model
- Extract a knowledge base from a short text

- Extract a knowledge base from a long text
- Filter and normalize entities

- Extract knowledge base of multiple entities from Wikipedia
- Visualize knowledge bases

Overall, by increasing the effectiveness and efficiency of question and answer systems, an automatic questions tagging system can offer considerable benefits for both users and companies.

III. Result



References

The paper ”T2kg: An end-to-end system for creating knowledge graph from unstructured text” by Kertkeidkachorn and Ichise describes a system for automatically constructing a knowledge graph from unstructured text data. The proposed system, called T2kg, consists of several modules that perform tasks such as named entity recognition, relation extraction, and entity linking to transform raw text into a structured knowledge graph [2].

The paper ”Accurate text-enhanced knowledge graph representation learning” by An et al. proposes a method for learning representations of entities and relations in a knowledge graph using both structured and unstructured data. The authors introduce a model called TEKGE (Text-Enhanced Knowledge Graph Embedding) that jointly optimizes over knowledge graph data and text data to learn entity and relation embeddings that capture both semantic and contextual information. The TEKGE model combines these two modules by defining a joint optimization objective that maximizes the likelihood of observed knowledge graph triples and text descriptions. The authors evaluate their model on two benchmark datasets and show that it outperforms several baseline methods that only consider either knowledge graph or text data.

Overall, [1] the paper presents a novel method for learning knowledge graph representations that combines information from both structured and unstructured data sources. The authors’ experiments demonstrate the effectiveness of their approach and suggest that incorporating text data can lead to more accurate and comprehensive knowledge graph representations. .

References

[1] An, B., Chen, B., Han, X. and Sun, L., 2018, June. Accurate text-enhanced knowledge graph representation learning. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 745-755).

[2] Kertkeidkachorn, N. and Ichise, R., 2017, March. T2kg: An end-to-end system for creating knowledge graph from unstructured text. In Workshops at the Thirty-First AAAI Conference on Artificial Intelligence.