Project 18: Construction of Knowledge Graph(KG) for IIT Kanpur website

Nitesh Trivedi[1], **Yash Kumar**[2], **Vishal Singh**[3]
[1]Roll No. 19211266, [2]Roll No. 19211274, [3]Roll No. 19211273
[1]CSE, [2]CSE, [3]CSE
{nitesht, yashk}@iitk.ac.in, vshlsng@cse.iitk.ac.in

Abstract

An unstructured text contains valuable information but retrieving elements of interest from the unstructured text requires crucial NLP techniques to process unstructured text. One such important element is the Knowledge Graph(KG). Most of the modern applications prepare knowledge base using KG and derive hidden insights. We aim to develop KG for the IIT Kanpur website using classical NLP techniques.

1 Introduction

Knowledge Graph is a graph-based model that encrypts the relation between entities in unstructured text. One of the examples of KG is DBpedia(Auer et al., 2007). Recently the trend of publishing unstructured data has grown over publishing structured data(Kríž et al., 2014). A renounced example is Google, it prepares KG using information gathered via a variety of sources and uses it to enhance search engine results. Many approaches discuss transforming unstructured text to structured text in order to develop KG (Carlson et al., 2010)(Exner and Nugues, 2012). All these approaches do very well on triples(subject, predicate, object) extraction but lack when it comes to mapping triples to the identical predicate in the KG (Kertkeidkachorn and Ichise, 2017). We propose an efficient way to develop KG for the IIT Kanpur website using statistical NLP and rule-based techniques. Similarity metrics are to be used for identical predicate matching from unstructured text to KG.

2 Problem Definition

We propose an efficient way to develop KG for the IIT Kanpur website. This involves end to end KG construction. From pre-processing of unstructured text to the generation of KG, various concepts are interlinked like Entity-Mapping, Co-reference Resolution, Triple Extraction, and Triple Integration.

3 Related Work

KG construction generally considers these three tasks: 1) Coreference Resolution, 2) Triplet Extraction and 3) Entity Linking. Based on these tasks, previous approaches can be divided as follows:

The first group (Carlson et al., 2010; Fader et al., 2011; Schmitz et al., 2012) focuses on knowledge extraction from unstructured text. NELL (Carlson et al., 2010) for triplet extraction, bootstrap constraints are used for learning new constraints. ReVerb (Fader et al., 2011) and OLLIE (Schmitz et al., 2012) are open information systems. Both these systems use syntactic and lexical patterns for triplet extraction. Even though these approaches extract triples from unstructured text, they do not consider entity mapping which as a result may cause entity ambiguity.

The second group (Cattoni et al., 2012; Augenstein et al., 2012; Kríž et al., 2014) worked on entity mapping and knowledge extraction. In studies mentioned in (Cattoni et al., 2012; Kríž et al., 2014), triplets were extracted from unstructured text using Natural Language Processing (NLP) techniques and then they were stored as RDF triples by using their own ontology. LODifier (Augenstein et al., 2012) used a named entity recognition system to extract a triplet and generates an RDF triple using WordNet representation without considering other ontologies. This approach solved the problem of entity ambiguity, the extracted triples were not integrated into other KGs.

For Entity Linking, we used survey paper and after going through all the techniques we decided to use the rule-based method for optimization of the knowledge graph constructed. The techniques we studied are mentioned in the (Shen et al., 2014). For Coreference Resolution we used Stanford CoreNLP which we studied from the research paper (Raghunathan et al., 2010). And the third major task for our project triplet extraction, we used Stanford Parser and Spacy for implementation. Stanford Parser technique, OpenNLP, Link Parser, and Minipar Parser studied from (Rusu et al., 2007).

4 Proposed Approach

We divide our work into three major modules 1) Entity-Mapping 2) Co-reference Resolution 3) Triple Extraction. Entity Mapping links the entity of unstructured text to the corresponding entity in KG. Co-reference resolution is performed to resolve the co-reference chain present in unstructured text. Triple extraction extracts triple from the text in the form of (Subject, Predicate, Object). After integrating results obtained by all modules, we aim to perform predicate linking using similarity metrics. Figure 1 explains our pipeline with the help of a simple example. Initially, we plan to apply our technique to the IIT Kanpur related text available on Wikipedia. Further, we intend to extend our work to test the accuracy of our model on the IIT Kanpur official website.

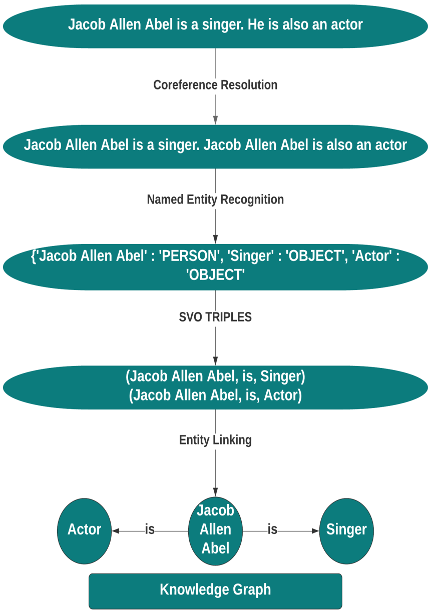
Jacob Allen Abel is a singer. He is also an actor
Coreference Resolution
Jacob Allen Abel is a singer. Jacob Allen Abel is also an actor
Named Entity Recognition
{'Jacob Allen Abel' : 'PERSON', 'Singer' : 'OBJECT', 'Actor' : 'OBJECT'
SVO TRIPLES
(Jacob Allen Abel, is, Singer)
(Jacob Allen Abel, is, Actor)
Entity Linking
Actor
is
Jacob Allen Abel
is
Singer
Knowledge Graph

Figure 1: Knowledge Graph Construction Pipeline

5 Corpus/Data Description

We are in process of scraping the IIT KANPUR website. After scraping we aim to preprocess scraped data. Further preprocessed corpora is mined to extract knowledge graph.

6 Experiments and Results

We considered a Sample Dataset(ActorsDataSet) from Wikipedia, and we ran the dataset via our KG construction pipeline. The results after preprocessing and coreference resolution were further processed to extract triplets. These triplets can further be improved as they are dependent upon the data which is the next stage of our work. The constructed Knowledge Graph is presented in 2. From the implementation perspective, we considered the two most widely used libraries spacy and Stanford. Comparing the results once we run our pipeline on our scraped corpus, we will compare the two and consider the one giving optimized results.

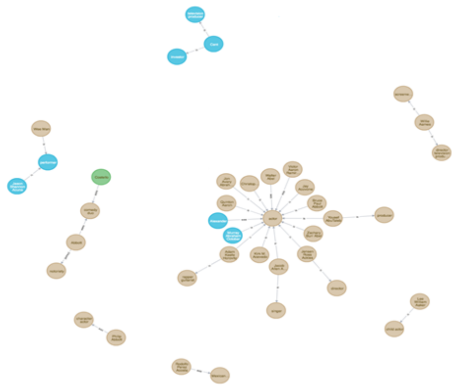

Figure 2: KG generated from toy data

7 Individual Contribution

Table 1: Contribution

Name	Contribution
Nitesh Trivedi	Webscraping and Coreference Resolution
Yash Kumar	Webscraping, NER and Triplet Extraction
Vishal Singh	Entity Linking

8 Future Work

We are in the process of scraping the IIT KANPUR website data and to pass the same through our knowledge graph construction pipeline. Once our knowledge graph gets constructed, we will query it for information retrieval. Retrieved results will be compared with the Google knowledge graph API to evaluate the performance.

9 Conclusion

Our knowledge graph construction primary pipeline is ready. We aim to pass our IIT KANPUR corpora through our pipeline and tune our modular structure. Due to the high dependency of the pipeline on data we have presented the knowledge graph obtained from toy data.

10 Presentation Feedback

One most important feedback we got is to carefully use the rule-based approach. Rule-based approaches may lead to unexpected results. Another suggestion was for Named Entity Recognition(NER) for which we got the suggestion to use labeled data and training the model based on the Probabilistic method and use the obtained model for Entity Recognition.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Isabelle Augenstein, Sebastian Padó, and Sebastian Rudolph. 2012. Lodifier: Generating linked data from unstructured text. In *Extended Semantic Web Conference*, pages 210–224. Springer.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Roldano Cattoni, Francesco Corcoglioniti, Christian Girardi, Bernardo Magnini, Luciano Serafini, and Roberto Zanoli. 2012. The knowledgestore: an entity-based storage system. In *LREC*, pages 3639–3646. Citeseer.

Peter Exner and Pierre Nugues. 2012. Entity extraction: From unstructured text to dbpedia rdf triples. In *WoLE@ ISWC*, pages 58–69.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics.

Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2kg: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Vincent Kríž, Barbora Hladká, Martin Nečaskỳ, and Tomáš Knap. 2014. Data extraction using nlp techniques and its transformation to linked data. In *Mexican International Conference on Artificial Intelligence*, pages 113–124. Springer.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multipass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Delia Rusu, Lorand Dali, Blaz Fortuna, Marko Grobelnik, and Dunja Mladenic. 2007. Triplet extraction from sentences. In *Proceedings of the 10th International Multiconference" Information Society-IS*, pages 8–12.

Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 523–534. Association for Computational Linguistics.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. volume 27, pages 443–460. IEEE.