International Institute of Information Technology, Hyderabad

CS4.501 Social Computing Spring-2021

Course Project - Second deliverable

Problem Statement and Scope Document

Enhance Knowledge Graphs/Knowledge bases like Wikidata

Team Members:
Pushpa Yadhav (2019900034)
Vijaya Lakhsmi (2018900071)
Karthikeyan Arumugam(2018900074)

Mentor:
Tushar Abhishek

February 15,2021

1 Introduction

Human knowledge provides a formal understanding of the world. Knowledge graphs that represent structural relations between entities have become an increasingly popular research direction towards cognition and human-level intelligence. Some of widely used Knowledge graphs in NLP related research are Freebase, YAGO and WikiData. In this project we would be exploring ways to make wikidata denser.

2 Problem statement

The amount of facts present in Knowledge graphs (KG) does not capture most of the world knowledge which is present in the text.The Quality of the Knowledge graph can be illustrated by density of edges (relations) and volume of information captured in it. There are various possibilities to enrich and improve the quality of the KGs.
Some of possible challenges at present are

- Knowledge graphs contains entities but not all relations are captured
- Entities are not captured in Knowledge graphs hence information pertaining to those entities are missing
- Augmenting different KGs/Sources to create enrich KG is tedious task.

3 Scope for this project work

- We will be using WikiData as our Base Knowledge Graph
- In this project, we will be focusing on only one specific Domain among our shortlisted domains
- We will be using Wikipedia pages as source to enrich domain we choose
- We will be using BERT based Pre-trained Language models as Input for the model

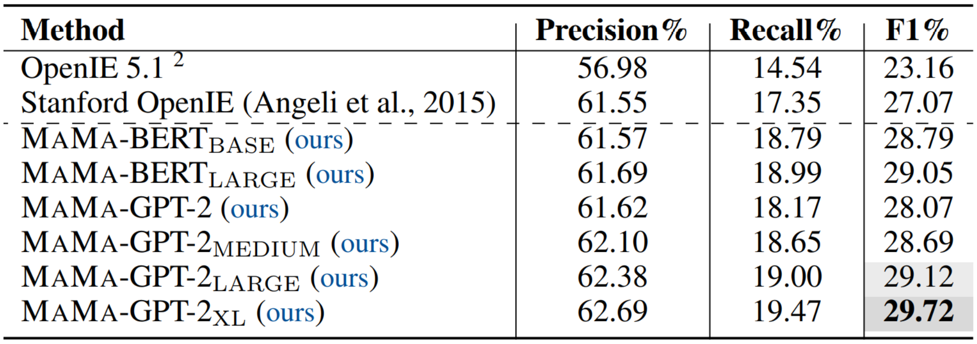
4 Proposed Project work and its end goals

We have decided to attempt approaches proposed in ”Language models are open knowledge graphs” by Chenguang wang et al,.2020.

1. We will be working only in ”English” language version of wikidata/ wikipedia pages. any other languages are out of our scope at this point.

2. After Considering wikidata stats page (https://www.wikidata.org/wiki/Wikidata:Statistics), we have short listed following domains for this project.
    - Film (Current Size - 294,370 or 0.4% of Wiki Data)
    - Chemical compound (Current Size - 1,188,724 or 1.7% of Wiki Data)
    - Through fares (Current Size - 630,794 or 0.9% of Wiki Data)
    - Wiki News Articles (Current Size - 195,900 or 0.3% of Wiki Data))
3. We would like to identify any issues and explore possible solution for such shortcomings of this model.
4. With help of our experiments we would investigate and report the accuracy measures.
5. Goal is to attempt to generate more Entries than what we have currently in wiki data. and compare how much amount of enrichment it brings to wiki data. In terms of number of edges (relations density) and number of nodes (Volume of Entities Captured)

4.1 Baseline (Part of Feedback from 1st Submission)



Method	**Precision%**	**Recall%**	**F1%**
OpenIE 5.1[2]	56.98	14.54	23.16
Stanford OpenIE (Angeli et al., 2015)	61.55	17.35	27.07
MaMa-BERT$_{\text{BASE}}$ (ours)	61.57	18.79	28.79
MaMa-BERT$_{\text{LARGE}}$ (ours)	61.69	18.99	29.05
MaMa-GPT-2 (ours)	61.62	18.17	28.07
MaMa-GPT-2$_{\text{MEDIUM}}$ (ours)	62.10	18.65	28.69
MaMa-GPT-2$_{\text{LARGE}}$ (ours)	62.38	19.00	29.12
MaMa-GPT-2$_{\text{XL}}$ (ours)	62.69	19.47	**29.72**

Table 1: Baseline Performance bench mark from existing model mentioned in original paper

We will be using above bench mark as our baseline model for this project, our project outcomes will be evaluated against this benchmark scores.

4.2 Proposed changes from Original paper

1. In addition to Standford OpenIE, we will also be using (REL) [https://github.com/informagi/REL] for entity disambiguation model (supervised instead of the original unsupervied) to achieve the same task and compare the results.

2. Having access to a relevant entity linker can save a lot of time and effort by identifying the right entities in the target KG. It plays a key role in being able to extract as much automatically, leaving less work for humans that would otherwise need to correct entities and relations

3. Tuning the thresholds for triplets with trial and error methodology.

4. The head and the tail entities are always single words, whereas many entities such as names have two or more words. This is one of the major critisim passed on about the original paper. In such cases we will try fine tune proposed model in original paper to capture more accurate entity names rather than just first/last words.

5 Literature Review summary

- **A Survey on Knowledge Graphs:Representation, Acquisition and Applications , Ji et al.,2021**

  This research conducted a comprehensive survey on 4 scopes

  1. knowledge graph embedding
  2. knowledge acquisition of entity discovery
  3. temporal knowledge graph representation learning
  4. real-world knowledge-aware applications

  This paper discussed deeper into theory and methods involves in various aspects of Knowledge graphs , they covered technical aspects such as knowledge representation, techniques to acquisition of new knowledge , various models and their comparison in detail. K-BERT,LSTM-CNN, RNN, BiLSTM,reinforcement learnings,and various other methods were discussed in detail and how it can be used in various stages of Knowledge graph lifecycle. This gave us lot of insight in state-of-art research effort going in Knowledge graph related area.

- **Accurate Text-Enhanced Knowledge Graph Representation Learning, Bo An et al,. 2018**

  This paper proposes representation frame work called "an accurate text enhanced knowledge graph representation framework", which enhance the knowledge representations of a triple, and effectively handle the ambiguity of relations and entities through a mutual attention model between relation mentions and entity descriptions. Their experiment results shows that their method can achieve the state-of-the-art performance, and significantly outperforms previous text-enhanced knowledge representation models.

- **Learning to Update Knowledge Graphs from Reading News, (Tang et al., 2019)** This research work proposes a novel graph based neural network method called GUpdater. GUpdater is build upon graph neural network (GNN) with a text based attention model. This model was able to effectively perform link-adding or link-deleting operations to ensure the KG up-to-date according to news snippets. Experiments demonstrated that this model can handle explicit and implicit information found in news sources.

- **Collective Multi-type Entity Alignment Between Knowledge Graphs, Qi Zhu1 et al.,2020**

  This research paper presents a new method ”Collective Graph neural network for multi-tyoe entity alignment” (CG-MuAlign). This method jointly aligns multiple types of entities, collectively leverages neighborhood information and generalizes to unlabled entite types. This experiment propose novel collective aggregation function tailored for reliving the incompleteness of the knowledge graphs via both cross-graph and self attentions, it also scales up effectively with mini-batch training paradigm and effective neighborhood sampling strategy. Their experiments with real world knowledge graphs with millions of enitites and they observed superior performance beyond exisiting methods. Running time of this method is much less that current state-of-the-art deep learning methods. Experiments demostrated that this proposed method can handle multiple knoeledge graphs alignment simultaneously.

- **Language Models are Open Knowledge Graphs, Chenguang Wang et al.,2020**

  This paper propose an unsupervised method to cast the knowledge contained within language models into KGs.Specifically it shows how to construct knowledge graphs (KGs) from pre-trained language models (e.g., BERT, GPT-2/3), without human supervision. This paper introduces a two-stage unsupervised approach called MaMa (Match and Mapping), which can successfully recover the factual knowledge stored in language models to build KGs from scratch. This MaMa Constructs a KG with a single forward pass of pretrained language models over a textual corpus. Further experiments with this model demonstrated that open knowledge graph features new facts when compared to WikiData and TAC KBP. This Model establishes a bridge between the deep learning and knowledge graphs, its results suggests that larger language models store richer knowledge than existing knowledge graphs

6 References and resources

- A Survey on Knowledge Graphs: Representation, Acquisition and Applications , Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, Philip S. Yu, 2020 , arXiv:2002.00388 [cs.CL]

- ”Accurate Text-Enhanced Knowledge Graph Representation Learning”, Bo An, Bo Chen, Xianpei Han and Le Sun,2018, ”Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 , *https://www.aclweb.org/anthology/N18-1068,10.18653/v1/N18-1068*

- Learning to Update Knowledge Graphs from Reading News, Jizhi Tang, Yansong Feng, Dongyan Zhao , Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) *https://www.aclweb.org/anthology/D19-1265, 10.18653/v1/D19-1265*

- Collective Multi-type Entity Alignment Between Knowledge Graphs, Qi Zhu , Hao Wei, Bunyamin Sisman, Da Zheng , Christos Faloutsos, Xin Luna Dong , Jiawei Han,2020, Proceedings of The Web Conference 2020, *https://assets.amazon.science/ff/7a/b962829844a0fbe5e31a8fcf68d17/scipub-1202.pdf, https://doi.org/10.1145/3366423.3380289*

- Where do Mayors Come From: Querying Wikidata with Python and SPARQL *https://janakiev.com/blog/wikidata-mayors/*

- Language Models are Open Knowledge Graphs , Chenguang Wang, Xiao Liu, Dawn Song , 2020, *arXiv:2010.11967*

- *https://query.wikidata.org/*

- Python Library for accessing Wikipedia pages*https://pypi.org/project/wikipedia/*

- Python Library for accessing Wiki Data*https://pypi.org/project/Wikidata/*

- Latest BERT Based Language model developed by Microsoft - *https://github.com/microsoft/DeBERTa*

- *https://github.com/google-research/bert*