Enhance Knowledge Graphs / Knowledge bases like Wikidata

CS4.501 Social Computing Spring-2021 - Project Team 9 - Course Project . Mentor : Tushar Abhishek (tushar.abhishek@research.iiit.ac.in)

Pushpa Yadhav
(2019900034)
IIIT-Hyderabad
pushpa.yadav@research.iiit.ac.in

Vijaya Lakhsmi
(2018900071)
IIIT-Hyderabad
vijaya.lakshmi@students.iiit.ac.in

Karthikeyan Arumugam
(2018900074)
IIIT-Hyderabad
karthikeyan.arumugam@students.iiit.ac.in

***Abstract*—The Quality of the Knowledge graph can be illustrated by density of edges (relations) and volume of information captured in it.The amount of facts present in Knowledge graphs (KG) may not capture most of the world knowledge and publicly available Knowledge graphs can be improved with more information.Pretrained Language models such as BERT are usually trained over large corpus and it inherently holds vast world knowledge. In this Project we are aim it to enrich the existing knowledge graphs using Pre trained knowledge graphs and wikipages corpus**

***Index Terms*—Wiki data, knowledge graphs, knowledge bases, language models, BERT,pre-trained language models, MAMA, Match and Map Algorithm**

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 course project we would be exploring ways to make wikidata denser by adding more Entities , extablishing more relations among entities etc.

LIERATURE SURVEY

Please refer all old submissions for detailed report on our literature survey. In this project we are attempting to implement and refine proposed models in Wang Et al. [6]

PROPOSED METHODS

In our proposed method we are using Wikidata as our base Knwoledge graphs, Source documents to mine Knowledge graph entries will be Wiki pages. We are focusing only on English wiki pages in this project

Source Documents

We used wikipediaapi to crawl through Wiki Pedia and we extracted page list to specific domain. All pages in such domains are extracted as text files and preprocessed before fed as Input to MAMA algorithm.

We downloaded pages for following domains for this experiments

- Films
- Chemical Compounds
- Through Fares
- News

Using Language models to match knowledge

We are using Pretrained Language models such BERT,GPT2 to identify the knowledge available in the source corpus. We will be using attention matrix generate from single forward pass with Pretrained LM , to match the entities and relations given in the Source documents

Entity Linking

Entity linking step is essential for us to make sure we are not duplicating the knowledge in existing KG. For entity linking we attempted to use following Systems

1) Stanford Core NLP - IE
2) REL
3) Spacy Entity Linker

EXPERIMENTS

Match and Map Algorithm - MAMA

Proposed model extras Knowledge from given corpus as triplets (head,relation,tail) using Match and map algorithm. This contains two step process, first part of the algorithm uses pretrained Language models to extract matched triplets and Second process validates and maps those triplets to existing knowledge graphs.

Match

The Match stage generates a set of candidate facts by matching the facts in the textual corpus with the knowledge in the pre-trained Language Model. Candidate facts are a set of extracted string triplets (namely; head, relation, and tail respectively). The longer the sentence the harder it is to extract the triplet strings.

We have run spaCy (spaCy is a free open-source library for Natural Language Processing in Python. It features NER,

POS tagging, dependency parsing, word vectors, and more) to extract noun phrases/chunks. Running the spaCy library through the corpus will find the head and tail of the extracted candidate facts. But the downfall to that spaCy is probably going to miss a lot of words that are not recognized as noun phrases (in huge sentences) and the authors also say that spaCy annotations are sometimes error-prone.

Mapping

Match process generated a csv file with subject , relation and objects. Mapping process used this input file to map it to WikiData. In our initial experiments , we used REL (supervised Model) for Entity linking with WikiData. We reused Pretrained model files shared as part of REL project (https://github.com/informagi/REL). Challenge with REL is the trained mode file is very huge and performance of REL model is not good enough. Entities resolved where returned as plain text title, instead of Entity ID in wikidata.

Subsequently we experimented with Standford Core NLP IE model

We have used standford CoreNLPClient for entity linking ,neuralcoref for coref resolution

In our final experiment, we used Spacy Entity linker (https://github.com/egerber/spacy-entity-linker) , this process simiplified and it was most effective in Linking given subject/object to an entity in WikiData.

Removing Invalid relations : often matched triplets may contain vague relation details. for example India is a country / India have states. Insuch case we have to process those relations and filterout invalid relations found in the list of Matched triplets.
We created listed of lexically invalid relations such as adverbs, adjectives etc. our mapping algorithm accounts for and removed triplets with such invalid relations.

CHALLENGES

Coreference Resolution in Input corpus

Coref resolution outputs are not perfect . for example "'Mary Frances Reynolds (April 1, 1932 – December 28, 2016), known professionally as Debbie Reynolds, was an American actress, singer, and businesswoman. Her career spanned almost 70 years." expected output "'Mary Frances Reynolds (April 1, 1932 – December 28, 2016), known professionally as Debbie Reynolds, was an American actress, singer, and businesswoman. Mary Frances Reynolds career spanned almost 70 years. but result from spacy neuralcoref is ['Mary Frances Reynolds (April 1, 1932 – December 28, 2016), known professionally as Debbie Reynolds, was an American actress, singer, and businesswoman.', 'an American actress, singer, and businesswoman career spanned almost 70 years.']

Mapping relations Using Entity resolver

With Current Entity Resolver models, we were able to resolve head and tail to entities in the existing Knowledge graphs i.e WikiData.
Tripplets generated by Match process of MAMA algorithm may contain invalid relation names such as adverbs , year etc. Such Triplets are not a valid KG entires hence we created a process to eliminate such invalid KG entries. Our Entity resolvers are not yet equipped to validate the existing relation/Edges defined in the Knowledge graphs. In our next iteration of Mapping Algorithm program, we are aiming to resolve this issue.
When edge IDs/relations cannot be found in the existing Knowledge graph data, we will be creating new entry using open knowledge graph notations instead

FUTURE STEPS

Source Documents

We will Scrape level 2 pages , current process extracted wiki pages which are level 1 in search for the given catergory/domain.

For Match

We need to further improve on coref resolution.

For Mapping

Entity linking to be improved to handle scenarios such as resolving relations to existing edge ID in wikidata

Pair wise Confidence Scoring

To evaluate the quality of knowledge extracted from the document , we will include pair wise confidence score.

ANALYSIS

Detailed analysis for the experiments and their resultant Knowledge graphs will be given as part of subsequent submissions.
This Analysis will include statistical data on how many new entities added , how many new relations extracted from source and detailed qualitative analysis on the experiment results.

RESOURCES

1) Assets and corpus - Google Drive folder
   a) http://tiny.cc/EnhanceWikiPrjData - Wiki Page Data
   b) http://tiny.cc/EnhanceWikiPrjIMPL -All Working Directory Files.
   c) http://tiny.cc/EnhanceWikiPrjAssets -REL Model files and other files used to run the program
2) https://github.com/Team9-CS4-501-S21-CourseProject/EnhanceWikiData - GitHub Repository

REFERENCES

[1] 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]

[2] "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*

[3] 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

[4] 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/b96282984a0fbe5e31a8fcf68d17/scipub-1202.pdf, https://doi.org/10.1145/3366423.3380289

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

[6] Language Models are Open Knowledge Graphs , Chenguang Wang, Xiao Liu, Dawn Song , 2020,
arXiv:2010.11967