International Institute of Information Technology, Hyderabad

CS4.501 Social Computing Spring-2021

Course Project - Second deliverable

Project Outline (final)

Enhance Knowledge Graphs/Knowledge bases like Wikidata

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

Mentor:
Tushar Abhishek

April 1,2021

1 Update Since Last Submission

- Quality of generated triplets from Match algorithm was not great in our last submission, Majorly due to coreference resolution challenges. Eg often in large documents, Subject reference which happens after first occurence will have third person/second person notation, hence generated Triplets often had Subject/Object entity as He , She, they etc.
- to over come this issue - we had to add additional preprocessing step to the existing clean up process to add coreference resolution to it.
- Coreference resolution issues in Pre Processing - 85 % completed
  - Hugging face - Core ref Approach - Model yields average performance, in many cases it is not resolving coreferences effectively and it is getting struck with one Object/Subject name and retains it for rest of the document
    eg

	subject	relation	object	su
0	American actress	released	American actress	N
1	American actress	include	The Singing Nun	N
2	Reynolds	reached	new younger generation	Pl
3	American actress	released	American actress	N
4	American actress	released	second autobiography	N
5	Reynolds	had	business ventures	Pl
6	Reynolds	received	the Screen Actors Guild Life Achievement Award	Pl
7	American actress	received	the Screen Actors Guild Life Achievement Award	N
8	Debbie	has	refreshing sense	Pl
9	Debbie	led to	role	Pl
10	Reynolds	reversed	Reynolds opinion	Pl
11	Debbie	portrayed	Jeanine Deckers	Pl
12	Debbie	played	Helen Chappel Hackett's mother	Pl
13	Debbie	played	recurring role	Pl
14	Debbie	had	cameo role	Pl
15	Reynolds	scored	hits	Pl
16	Debbie	released	album	Pl
17	production	broke	records	N
18	Debbie	received	Tony nomination	Pl
19	replacement	own	West End	N
20	2008	displayed	items	D
21	Reynolds	opened	Reynolds own dance studio	Pl
22	Debbie's Way!. Reynolds	purchased	the Clarion Hotel and Casino	Pl
23	The Eddie Fisher – Elizabeth Taylor affair	noted	bright	O

    Hence we didnt used hugging face approach any further.
  - AllenNLP - Coreference resolution yields better results than hugging face neural coref library, Challenge with size of the ip document , we ran into issues when input document is greater than 1000 characters long.
- Map - Algorithm updates
  - In Last submission we were able to provide entity linking, hence generate triplets were mapped to existing KG entities. but we had challenge with Resolving relationship/edges text to relation ID in wiki data

- We employed SPARQL query to collect list of preexisting relations in the Wiki data. used word similarity to identify if generated relationship/edge is preexisting entry in wikidata
- Relationship resolution / Edge resolution to existing entries in Wikidata : Partly completed.- 50% , Few more tweaking and changes are required to make it more functional
- Confidence Score for the generated Triplets are computated using P(generated triplets) in language model. More fine tuning is required to build a confidence score measure.
-

2 Work In progress Action Items

- Running the entire pipeline for large volume of data - 40% completed
- Qualitative evaluation - 20 %
- Metrics to measure increase in density of the KG - 20%
- relationship resolution using SPARQL query - Fine tuning
- confidence score of generated triplets - Fine tuning
- Qualitative evaluation for manually comparing a single wiki page - to gauge accuracy of mining KG graph entries from the proposed method.

3 Remaining Action Items and Target dates

- Running entire data pipeline for Already scrapped larse corpus - 4th April
- Qualitative Evaluation of the Sample generated KG - 6th April 2021
- Quantitative metric results - 8th April 2021
- Final Presentation and reports - 10th April 2021

4 References

Git Hub Link: https://github.com/Team9-CS4-501-S21-CourseProject/EnhanceWikiData/