Improving open domain question answering with Knowledge Base and Wikipedia graph

Ruiyu Lin
SUN YAT-SEN UNIVERSITY
linry23@mail2.sysu.edu.cn

ABSTRACT

A clear and well-documented LaTeX document is presented as an

KEYWORDS

neural networks

ACM Reference Format:

Ruiyu Lin. 2018. Improving open domain question answering with Knowledge Base and Wikipedia graph. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Open-domain Question Answering mostly focus on factoid question answering,which require systems to return a short and concise answer to these questions. Most existing models, however,answer questions using a single information source, usually either text from an text corpus such as Wikipedia[3],or a single knowledge base (KB).

Large-scale factual knowledge bases such as WikiData and Freebase[2], stores a large number of facts in an organized way. Each fact is made of two entities and a relation between them.Most konwledge bases are curated,ensureing the correctness of the information,common or "simple" questions can be answered easily if semantic parsing (question query) is done correctly.The advantage of graph structure also enables multi-hop question answering. Unfortunately, curated konwledge bases, which demands tremendous hunman labor,might not keep up with times,thereby some relations would be missing.Limited coverage of questions can be answered because the resoning is based on the similarity over relationships and entities.

Wikipedia[3], a text source, was proposed for the first time to process Open Domain QA tasks, and a DRQA system was developed, including Document Retriever and Document Reader, which laid the pipe-line, two-stage approaches, of QA for successive work.We also follow this tradition ,retrieve and then read. Text corpus provides a more completed coverage of facts,and it is easy to catch the time,however lacks the ability of multi-hop resoning.

To combine the coverage of text evidence and reasoning ability of knowledge base, some recent work use both text and KB,to constructs graphs of nodes and edges[5, 6, 8].These works basically augment the KB graph with the entities indentified from the relevent text evidences,the task of answer determination is then reduced to classfiy the entity node is the answer or not. Another line[4] ,inversely,augment retrieval passage with KB graph, and the task of answer determination is to do answer extraction from text.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2 RELATED WORK

2.1 QA using Text

2.2 QA using KB

2.3 QA using both Text and KB

In *GRAFT-Net*[6] , *PullNet*[5],and *Knowledge-Aware*[8],the answers is restricted to be the KB entities. Meanwhile, their question subgraph is heterogeneous, which contains KB triples(entity,relation,entity), and entity-linked text. The task of QA then reduces to learning the representations of the nodes, and then performing a binary classification over these nodes to decide whether it is the answer or not. They both augmented knowledge bases with text from Wikipedia, which means KB dominates the whole process.

However,*Knowledge guided*[4] construct the graph in a different way. Inversely, the knowledge base is used to better model relationships between different passages of text, which means the text corpus dominates instead. Its question subgraph is not heterogeneous, only contains entity-linked passage and relations. The task of QA switches to learning the representations of the passage . Not to classify the node is the answer or not, it extracts the most possible span as answer in the most possible passage as prior work did.Our work is consistent with it.

3 METHOD

Input

- Dataset:(question, answer) pair.
- Knowledge Base:(entity, relation, entity) triple. Knowledge Base is a multi-relational graphs, each edge has a label and direction associated with it, and each node in the graph is an entity.
- Wikipedia graph[1]:(passage, passage) pair. Wikipedia graph is constructed by hyperlinks and within-document links,the edge is direct ,and each node in the graph is an article.The Wikipedia graph is densely connected and covers a wide range of topics that provide useful evidence for open-domain questions

Output representation of all the retrieved passage, as the input to a reader model to extract answer.

Goal To better embed the retrieved passage, which based on Wikipedia graph, with Knowledge Base knowledge.

Method Fuse Knowledge Base knowledge into Wikipedia graph to formulate (passage, relation, passage) triple.

(1) get the seed passage by a TF-IDF based retrieval system.Around the seed passage,take the neighbouring passage from Wikipedia graph,$(P_1, ..., P_N)$
(2) identify the relations between each two passage node assuming that :
$rSet_{(a,b)}$: the set of relation between P_a and P_b, initialized empty
(P_a, P_b) exits in Wikipedia graph.
P_a contains n entities($e_{a1}, ..., e_{an}$) ,
P_b contains m entities($e_{b1}, ..., e_{bm}$)
If (e_{ai}, r, e_{bj})(1<=i,j<=n) exits in Knowledge Base, add r into $rSet_{(a,b)}$. Finally, $rSet_{a,b} = (r_1 ..., r_k)$
(3) corperate identified relations into passage node embedding. Suppose that h_b is link by $[h_{a1}, ..., h_{at}]$,then update h_b by

$$\alpha_{r_i} = sorce(h_{a_i}, e_{ai}, r_i, q)$$
$$h_b = GCN(h_b, \sum_{1}^{t} FNN(h_{a_i}, \sum_{i=1}^{k} \alpha_{r_i} * ri))$$

α_{r_i} is the relation score.
(4) answer extraction
Denote the passage score as $Pr(P_i|Q, P)$,which reranks all retrieved passages

$$Pr(P_i|Q, P) = softmax(h_i^{\mathrm{T}} W)$$

W is a trainable parameter.
The score of an answer span from passage P_i will be

$$Pr(a|Q, P) = Pr(P_i|Q, P)P_s(a_s|Q, P)P_e(a_e|Q, P).$$

Some details remains to be determined

(1) how to get the representation of relation.
In GRAFT-Nets[6], they average word vectors to compute a relation vector from the surface form of the relation.
In PullNet[5] ,embedding of relations are pretrained ,and can be looked up from an embedding table.
In [4] ,they only consider the most frequent 100 relations, and pretrain their embedding.
[8] tokenize the relation ,and then encode it by a shared LSTM with the question .
In [7],it handles multi-relational graphs representation where each edge has a label and direction associated with it, and jointly embeds both nodes and relations in a relational graph.
(2) the representaion of question
In GRAFT-Nets[6], the question representation is updated as $h_q = FFN(\sum_{v \in S_q} h_v)$, where S_q denotes the seed entities mentioned in the question.
In PullNet[5]and [8] , the question representation is accquired by a LSTM.
In [4] ,the question is not directly encoded,but with passage jointly by BERT.
(3) how to score relation attention
Most work take the dot product between the embedding of relation and question.Considering the different framework here,we reformulate the score as:

$$\alpha_{r_i} = sorce(h_a, e_{ai}, r_i, q)$$

REFERENCES

[1] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. ICLR 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. (ICLR 2020).
[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 1247–1250.
[3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051* (2017).
[4] Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Knowledge guided text retrieval and reading for open domain question answering. *arXiv preprint arXiv:1911.03868* (2019).
[5] Haitian Sun, Tania Bedrax-Weiss, and William W Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2380–2390,7.*
[6] Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4231–4242.*
[7] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar. ICLR 2020. Composition-based Multi-Relational Graph Convolutional Networks. (ICLR 2020).
[8] Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving question answering over incomplete kbs with knowledge-aware reader. *arXiv preprint arXiv:1905.07098* (2019).

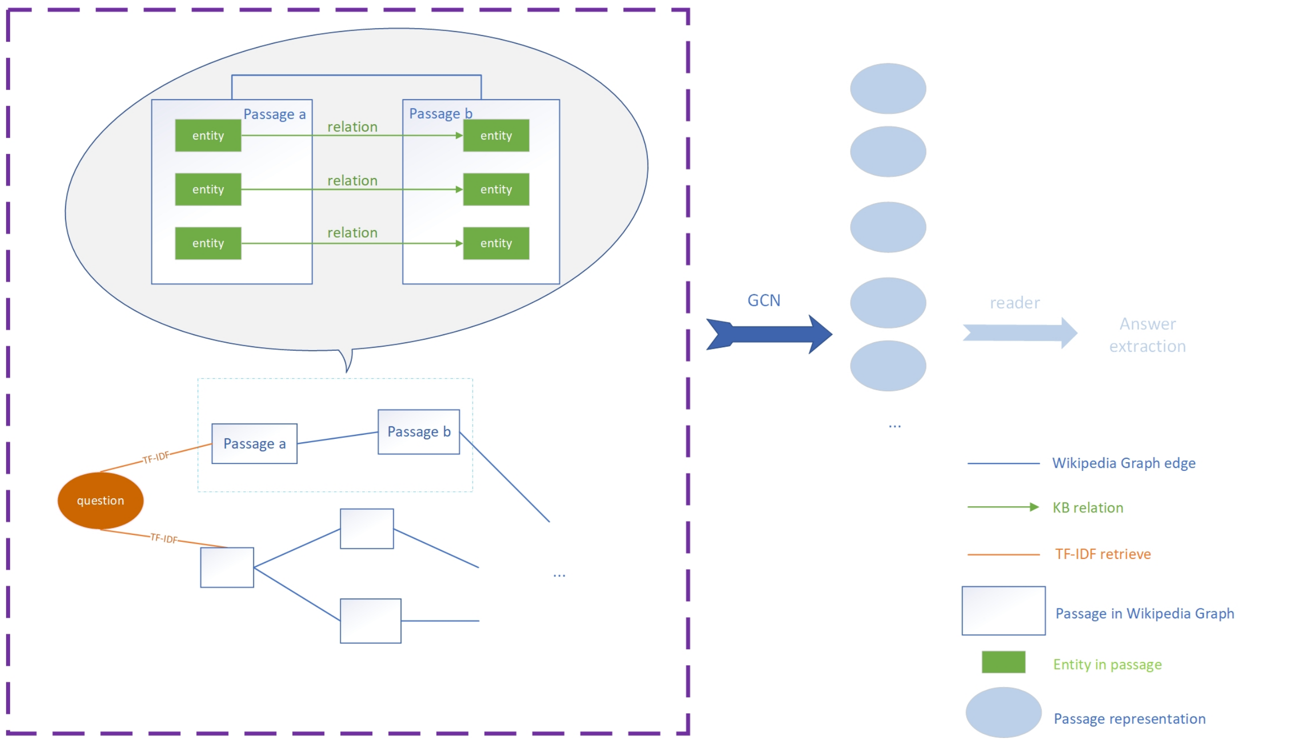

Figure 1: A diagram of approach