Improving open domain question answering with Knowledge Base and Wikipedia graph

ABSTRACT

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

KEYWORDS

neural networks

ACM Reference Format:

. 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[4],or a single knowledge base (KB).

Large-scale factual knowledge bases such as WikiData[11],Freebase [3],Dbpedia[2], stores a large number of facts in an organized way. Namely, Freebase has 46m entities and 2.6b facts, WikiData contains 87m items .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[4], 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 easier 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[9, 10, 12].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[5, 8] ,inversely,augment retrieval passage with KB graph, and the task of answer determination is to do answer extraction from text.

2 RELATED WORK

2.1 QA using Text

Cognitive Graph[5]retrieves evidence documents offline, and trains a reading comprehension model to jointly predict possible answer spans and next-hop spans to extend the reasoning chain. [1]learns to sequentially retrieve evidence paragraphs over the Wikipedia graph by conditioning on the previously retrieved documents.They constructs Wikipedia graph offline and reused it throughout training and inference for any question.We will use this new meta resource in our work.

2.2 QA using KB

2.3 QA using both Text and KB

In [9, 10, 12],the answers is restricted to be the KB entities. GRAFT-Net [10] constructs a heterogeneous question subgraph, which contains KB triples(entity,relation,entity), and entity-linked text. Afterwards, PullNet [9] expands more entities from relevant documents to form a isomorphic graph. 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,the text here plays an auxiliary role.

However,[8] construct the graph in a different way. Inversely, the knowledge base is used to better model relationships between different passages of text. It use KB relations to formulate entity-introductory-passage relations. The task of QA switches to learning the KB-aware representations of passages . 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 but with different approach.

3 METHOD

Specifically, we use Wikipedia graph[1] as the text corpus $\mathcal{G}$ and Wikidata[11] as the knowledge base $\mathcal{K}$, as there exists an alignment between the two resources. 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] is a directed graph,constructed by hyperlinks and within-document links, each node represents a single paragraph.The Wikipedia graph is densely connected and covers a wide range of topics that provide useful evidence for open-domain questions. Our goal is to fuse

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/10.1145/1122445.1122456

Knowledge Base knowledge into Wikipedia graph passages,learn the KB-aware representation of all the retrieved passage, as the input to a reader model to extract answer.

Seed Nodes Given a natural language question q,we use the top K_{TF-IDF} paragraphs returned by a TF-IDF based retrieval system as the start nodes.

Subgraph Next,we run Topic-sensitive pagerank[7] around these seed nodes to get the edges weight over their neighbouring, denoted as $\mathcal{G}_q = (\mathcal{P}, \mathcal{E})$,where $\mathcal{P}$ is the set of paragraphs $\{p_1, ..., p_{|p|}\}$ in the Wikipedia graph, and the edge $\mathcal{E}$ are links between them with a pagerank score pr over them.Each paragraphs is a sequence of words $p_i = (w_1^{p_i}, ..., w_{|pi|}^{p_i})$

Node Initialization Note that we use the same encoder BERT during the whole process.

(1) Question initialization
Given a natural language question $q = (w_1^q, ..., w_q^q)$,initialize question as:

$$h_q^{(0)} = BERT(w_1^q, ..., w_{|q|}^q)$$

We perform an entity linking system[6] to identity the *topic entity* in the question,denoted as $E_{topic} = \{e_1, ..., e_{|topic|}\}$.

(2) Passage initialization
And the passages in $\mathcal{G}_q$ are initialized as:

$$h_{p_i}^{(0)} = BERT(w_1^{p_i}, ..., w_{|pi|}^{p_i})$$

We denote the embedding of *pos-th* word in the passage p_i as $h_{pi}^{(l)}[pos]$.

Node Update Assuming that $(p_a, p_b) \in \mathcal{G}_q$, and the edge weight is $pr_{a,b}$, we also perform an entity linking system[6] to identity the entity in p_a and p_b.Suppose that p_a contains n entities ($e_{a1}, ..., e_{an}$) , p_b contains m entities ($e_{b1}, ..., e_{bm}$) .In each layer l,we update the information as follow:

(1) Corperate entity relations into passage .
We update passage representaion at token level. If the triple $(e_{ai}, r, e_{bj}) \in \mathcal{K}$, and e_{ai} is at the position pos_a in p_a, e_{bj} is at pos_b in p_b, let $L = (p_a, pos_a)$ be the set of all the entities appearing at position pos_a in paragrah p_a,they are all linked to the entities at position pos_b in paragrah p_b.We aggregate over the token hidden states coming in at each position separately:

$$h_{p_b}^{(l)}[pos_b] = FFN\left(\begin{bmatrix} h_{p_b}^{(l-1)}[pos_b], \\ \sum\limits_{(p_a,pos_a)\in L} pr_{a,b} * s(h_r, h_q^{(l)}) * \phi(h_{p_a}^{(l-1)}[pos_a], h_r) \end{bmatrix}\right)$$

$s(h_r, h_q^{(l)})$ is the relevant score between relation and question,we simply computes it through the dot product between their hidden state.h_r is embedding of relations,which are pretrained and can be looked up from an embedding table.

$$s(h_r, h_q^{(l)}) = h_r \dot{h}_q^{(l)}$$

ϕ is a transform operation,which is

$$\phi(h, r) = FFN(h; r)$$

(2) Update passage representation
Next we aggregate states within the paragraph using BERT again:

$$h_p^{(l)} = BERT(h_p^{(l)}[1], ..., h_p^{(l)}[|p|])$$

(3) Update topic entities representation
We have $E_{topic} = \{e_1, ..., e_{|topic|}\}$ from above step. Let $Q = (e, p, pos)$ denotes the topic entity e appear at the position pos in passage p,we update the topic entity representation by aggregate the information from the updated passages.

$$h_e^{(l)} = \frac{1}{|Q|} \sum_{(e,p,pos)\in Q} h_p^{(l)}[pos]$$

(4) Update question representation
Next we aggregate states within the qusetion using BERT again:

$$h_q^{(l)} = BERT(h_q^{(l)}[1], ..., h_q^{(l)}[pos])$$
$$s.t.$$
$$h_q^{(l)}[i] = \begin{cases} h_e^{(l)}, & \Phi(e, i) = 1 \\ h_q^{(l-1)}[i], & \Phi(e, i) = 0 \end{cases}$$

$\Phi(e, i) = 1$ means the token at position i in question is a topic entity.

Answer Extraction

Denote the passage score as $Pr(P_i|Q, P)$,which reranks all retrieved passages

$$Pr(P_i|Q, P) = softmax(W[h_{pi}^{\mathrm{T}}; h_q])$$

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

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

REFERENCES

[1] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *ICLR.*

[2] 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.* Springer, 722–735.

[3] 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.

[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051* (2017).

[5] Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. ACL 2019. Cognitive graph for multi-hop reading comprehension at scale. (ACL 2019).

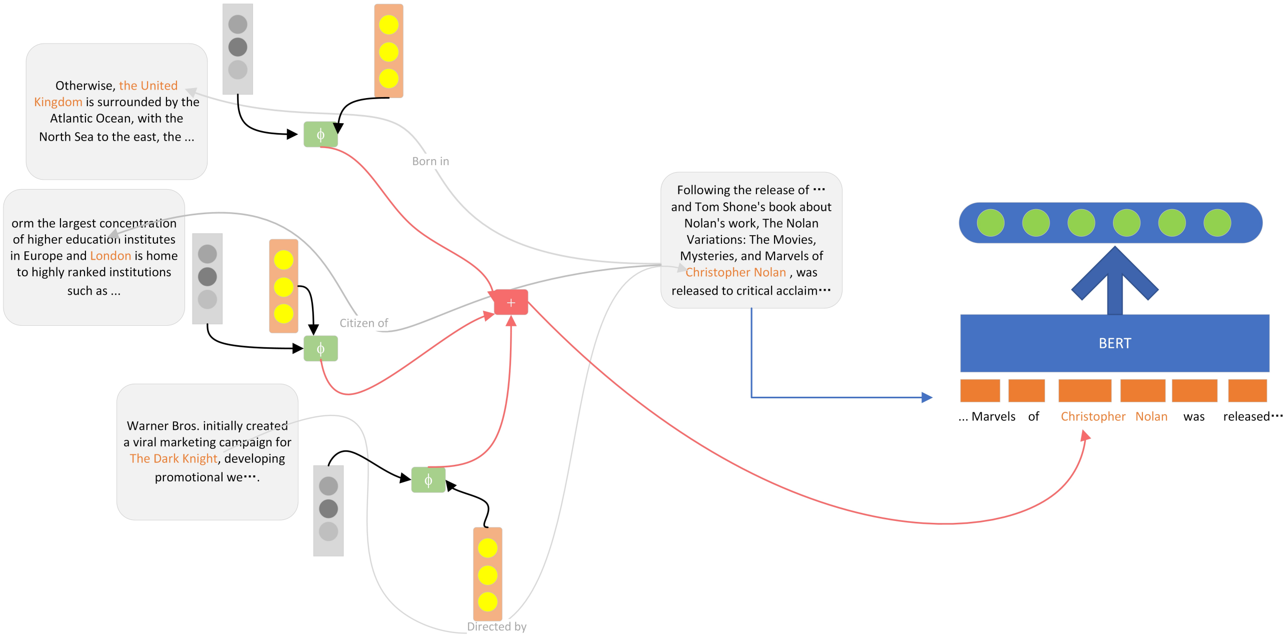
[6] Paolo Ferragina and Ugo Scaiella. 2011. Fast and accurate annotation of short texts with wikipedia pages. *IEEE software* 29, 1 (2011), 70–75.

[7] Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering* 15, 4 (2003), 784–796.

[8] 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).

[9] 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.*

[10] 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.*



Otherwise, the United Kingdom is surrounded by the Atlantic Ocean, with the North Sea to the east, the ...

orm the largest concentration of higher education institutes in Europe and London is home to highly ranked institutions such as ...

Warner Bros. initially created a viral marketing campaign for The Dark Knight, developing promotional we···.

Following the release of ··· and Tom Shone's book about Nolan's work, The Nolan Variations: The Movies, Mysteries, and Marvels of Christopher Nolan , was released to critical acclaim···

Figure 1: A diagram of approach

[11] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[12] 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).