Survey on Multi-hop Question Answering

Woojeong Jin
University of Southern California
woojeong.jin@usc.edu

Abstract

Learning multi-hop reasoning has been a key challenge for reading comprehension models. Ideally, a model should not be able to perform well on a multi-hop ques- tion answering task without doing multi-hop reasoning. In this survey, we investigate works on multi-hop question answering including datasets and models.

1 Introduction

Multi-hop question answering requires the aggregation of evidence across several paragraphs to answer a question. Table 1 shows an example of single-hop and multi-hop questions. A single-hop question "Which player is named 2015 Diamond Head Classic's MVP?" requires finding the player who won MVP from one paragraph. However, a multi-hop question requires further reasoning, which is first finding the player, and then finding the team that player plays for from another paragraph.

In this project, we explore and examines previous models for multi-hop question answering. We first study datasets for the multi-hop questions and then examine the literatures.

2 Dataset

Before diving into previous work, we first examine the datasets (Dua et al., 2019; Yang et al., 2018; Welbl et al., 2018; Talmor and Berant, 2018). Among them, we focus on two datasets: WikiHop (Welbl et al., 2018) and HotpotQA (Yang et al., 2018). One key difference is that HotpotQA is span-based (the answer is a span of the passage) while WikiHop is multiple-choice.

WikiHop. Wikihop is English dataset designed for text understanding across multiple documents. The dataset consists of 40k+ questions, answers, and passages, where each passage consists of several documents collected from Wikipedia. Questions are posed as a query of a relation r followed by a head entity h, with the task being to find the tail entity t from a set of entity candidates E. Annotators followed links between documents and were required to use multiple documents to get the answer.

Single-hop	Which player is named 2015 Diamond Head Classic's MVP?
Multi-hop	Which team does the player named 2015 Diamond Head Classic's MVP play for?

Table 1: An example of single-hop and multi-hop questions from HotpotQA. A multi-hop question requires multi-hop reasoning.

HotpotQA. HotpotQA is a dataset with 113k English Wikipedia-based question-answer pairs. The questions are diverse, falling into several categories: inferring the bridge entity, intersection, and comparison. All require finding and reasoning over multiple supporting documents to answer. Models should choose answers by selecting variable-length spans from the documents. Sentences relevant to finding the answer are annotated in the dataset as "supporting facts" so models can use these at training time as well.

3 Previous Work

In this section, we review the previous work on HotpotQA (Min et al., 2019; Xiao et al., 2019; Nishida et al., 2019; Ding et al., 2019; Feldman and El-Yaniv, 2019). Before digging into each method, Multi-hop reading comprehension has two benchmark settings on HotpotQA: distractor and full wiki (open-domain) setting. In the first setting, to challenge the model to find the true supporting facts in the presence of noice,

there are 8 paragraphs from Wikipedia as distractors, and 2 gold paragraphs, which contain answers and supporting facts). The second setting truly test the model's ability to locate relevant facts as well as reasoning about them by requiring it to answer the question given the first paragraphs of all Wikipedia articles without gold paragraphs specified.

We first divide the previous work into two groups: models for the distractor setting (Xiao et al., 2019; Min et al., 2019; Nishida et al., 2019), models for the full wiki (open-domain) setting (Min et al., 2019; Ding et al., 2019; Feldman and El-Yaniv, 2019).

Xiao et al. (2019) proposed the Dynamically Fused Graph Network (DFGN). Their intuition is drawn from the human reasoning process for QA. One starts from an entity of interest in the query, focuses on the words surrounding the start entities, connects to some related entity either found in the neighborhood or linke by the same surface mention, repeats the step to form a reasoning chain, and lands on some entity or snippets likely to be the answer. More specifically, they first find a paragraph and construct an entity graph. From these paragraph and graph, they find an answer.

Min et al. (2019) proposed DecompRC that learns to break compositional multi-hop questions into simpler, single-hop sub-questions using spans from the original question. First, DecompRC decomposes the original, multi-hop question into several single-hop sub-questions according to a few reasoning types in parallel, based on span predictions. Then, for every reasoning types DecompRC leverages a single-hop reading comprehension model to answer each sub-question, and combines the answers according to the reasoning type. Finally, it leverages a decomposition scorer to judge which decomposition is the most suitable, and outputs the answer from that decomposition as the final answer.

Nishida et al. (2019) proposed Query Focused Extractor (QFE) model for evidence extraction. So they focus on the evidence extraction based on the previous work (Yang et al., 2018). QFE is inspired by extractive summarization models; compared with the existing method, which extracts each evidence sentence independently, it sequentially extracts evidence sentences by using an RNN with an attention mechanism on the question sentence. It enables QFE to consider the dependency among the evidence sentences and cover important information in the question sentence.

The following work is for the full wiki setting. Thus, finding relevant paragraphs is crucial for their performances.

Ding et al. (2019) proposed Cognitive Graph QA (CogQA), which comprises System 1 and 2 modules. System 1 extracts question-relevant entities and answer candidates from paragraphs and encodes their semantic information. Extracted entities are organized as a cognitive graph. System 2 conducts the reasoning procedure over the graph, and collects clues to guide System 1 to better extract next-hop entities. The above process is iterated until all possible answers are found, and then the final answer is chosen based on reasoning results from System 2.

Feldman and El-Yaniv (2019) proposed MUPPET (multi-hop paragraph retrieval) which relies on the following basic scheme consisting of two main components: (a) a paragraph and question encoder, and (b) a paragraph reader. The encoder is trained to encode paragraphs into d-dimensional vectors, and to encode questions into search vectors in the same vector space. Then a maximum inner product search algorithm is applied to find the most similar paragraphs to a given question. The most similar paragraphs are then passed to the paragraph reader, and extracts the most probable answer to the question.

The above work showed good performances, but they are out of date. We can find the better models on the leaderboard of HotpotQA (`https://hotpotqa.github.io`).

References

Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive graph for multi-hop reading comprehension at scale. *arXiv preprint arXiv:1905.05460*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. *arXiv preprint arXiv:1906.06606*.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. *arXiv preprint arXiv:1906.02916*.

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction. *arXiv preprint arXiv:1905.08511*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Yunxuan Xiao, Yanru Qu, Lin Qiu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. *arXiv preprint arXiv:1905.06933*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.