LREC-COLING 2024

TOWARDS GRAPH-HOP RETRIEVAL AND REASONING IN COMPLEX QUESTION ANSWERING OVER TEXTUAL DATABASE

Minjun Zhu[1,2], **Yixuan Weng** [1], **Shizhu He**[1,2,✉],
Kang Liu[1,2,4], **Haifeng Liu**[3], **Yang Jun**[3], **Jun Zhao**[1,2]
[1] The Laboratory of Cognition and Decision Intelligence for Complex Systems, IA, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Guangdong OPPO Mobile Telecommunications Corp.,Ltd.
[4] Shanghai Artificial Intelligence Laboratory
minjun.zhu23@gmail.com, {shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn

LREC-COLING 2024 Poster

Introduction

Background and Importance:

Textual question answering (TQA) systems are crucial for extracting and reasoning over knowledge embedded in unstructured texts. They facilitate a wide range of applications from academic research to industry solutions.

Research Question/Objectives:

Our research aims to introduce a novel 'Graph-Hop' paradigm that enables more fine-grained and adaptable retrieval and reasoning in complex question answering over textual databases.

Innovation:

We propose the first benchmark, ReasonGraphQA, which provides complex structured retrieval assistance through interpretable evidence graphs.

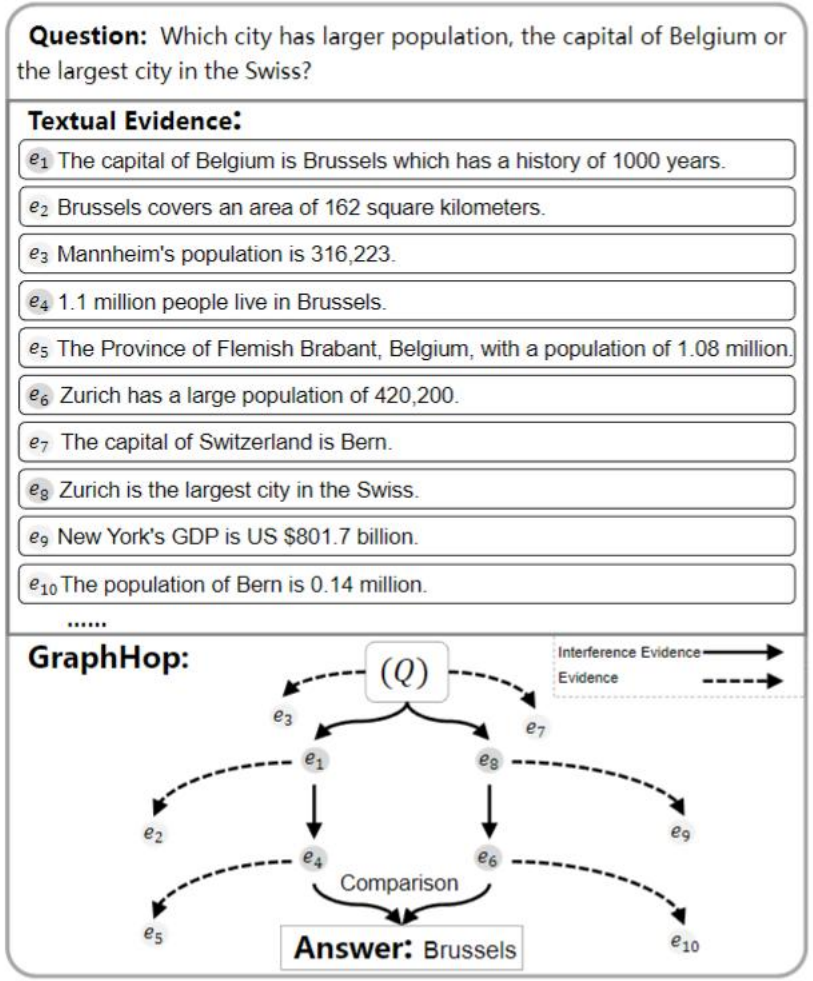
Graph-Hop

ReasonGraphQA is an advanced method used for querying and reasoning over structured data represented in graph formats. It leverages connections between data points to deduce logical conclusions.

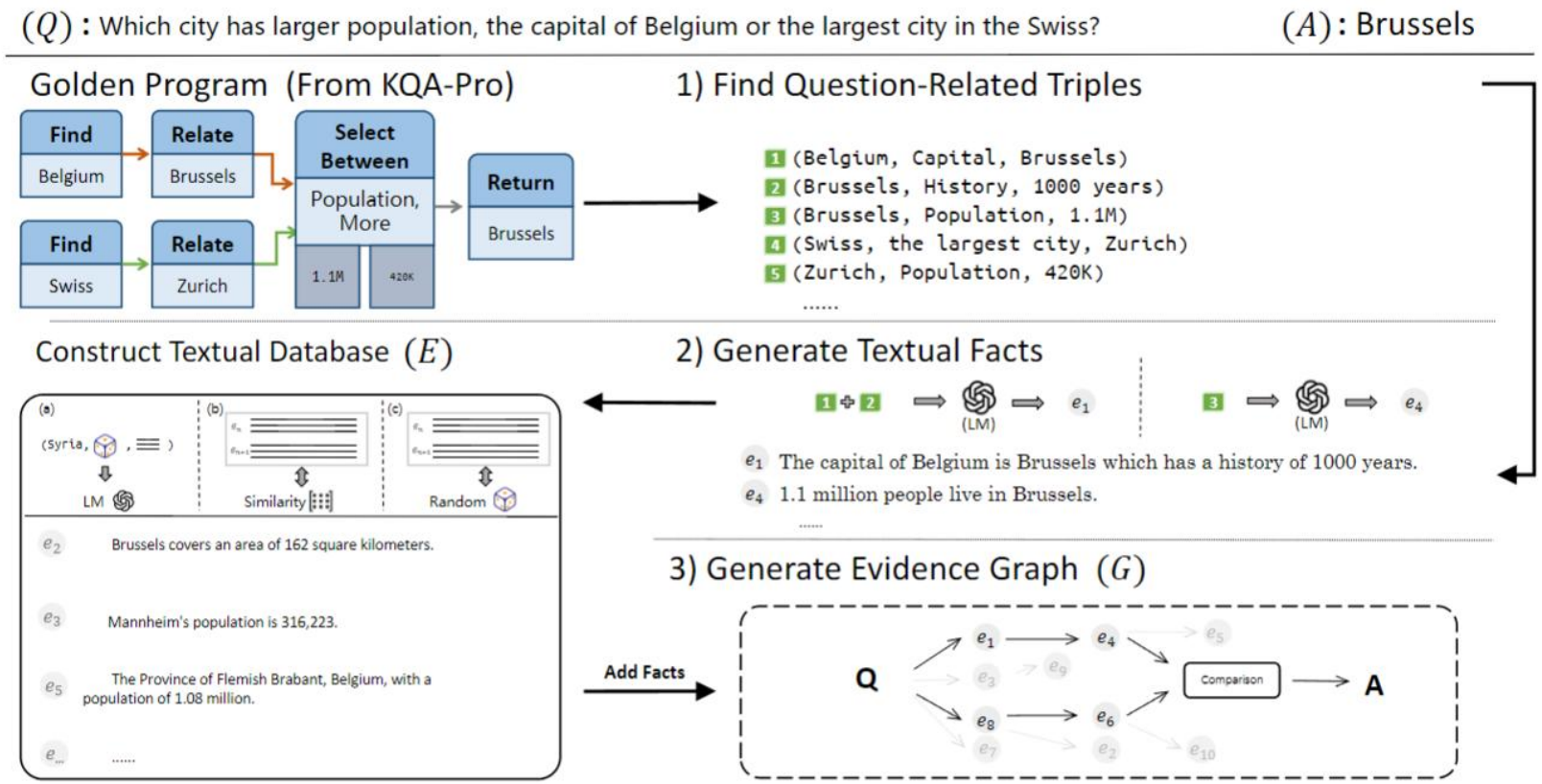
Graph-Hop Visualization:

The displayed graph visualizes how different pieces of evidence (e1, e4, e6, e7, e10) are interconnected to deduce the answer.
This approach compares populations using direct data (e.g., e4 and e10) and through inference across multiple nodes (e.g., e6 connected through e7 to e10).

Based on the evidence and logical paths depicted in the graph, Brussels has a larger population than both major Swiss cities combined, making it the clear answer to the question posed.



ReasonGraphQA Dataset Construction



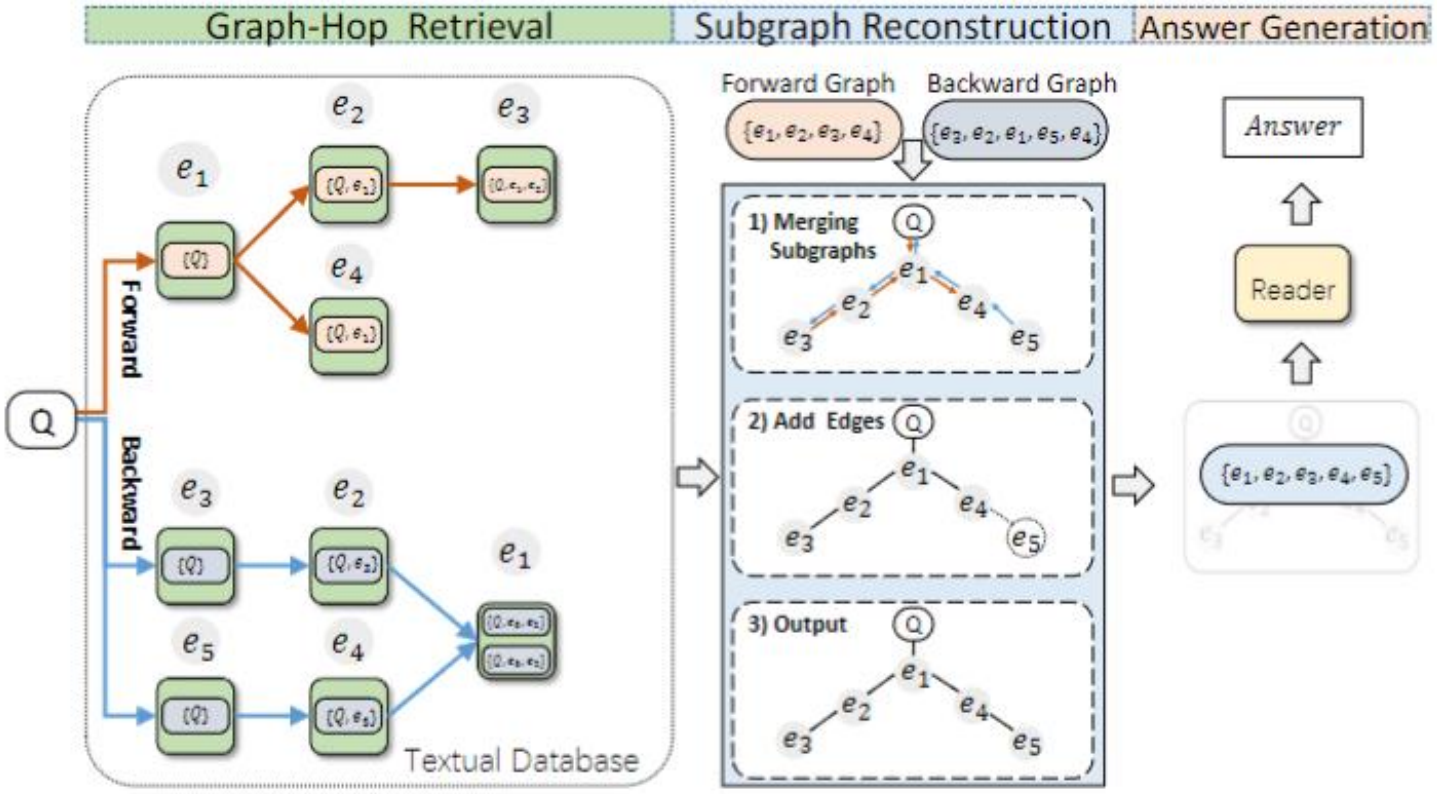
Developed a new benchmark called ReasonGraphQA which provides a structured approach for evaluating complex question answering systems.

The dataset features:

Questions requiring multi-hop reasoning across multiple evidence chains.

Fine-grained evidence graphs detailing the reasoning process.

Bidirectional Graph Retrieval



Combines both forward and backward retrieval processes to enhance reasoning accuracy and evidence recovery.

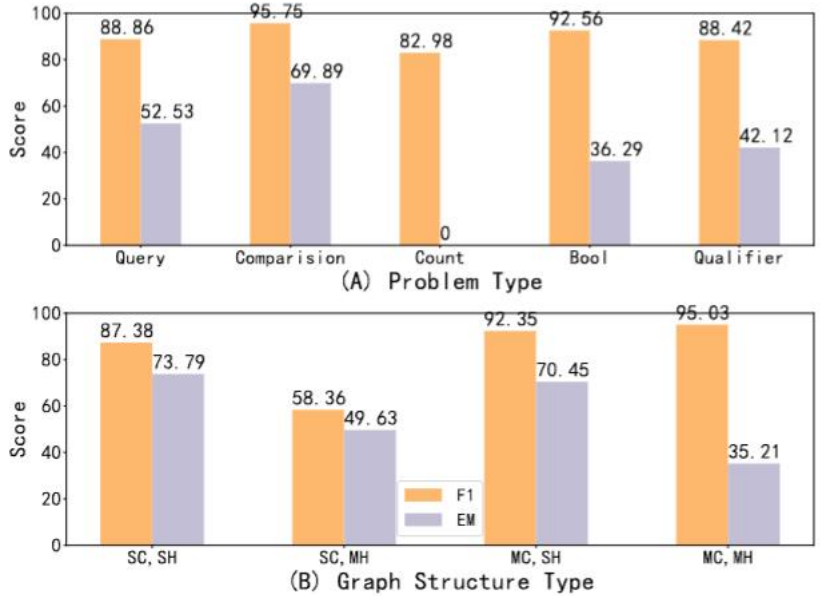
Evaluation and Results

Method		Explanation Graph			Evidence Set				QA EM
		GM↑	GS↑	GED↓	F1↑	Precision↑	Recall ↑	EM↑	Acc↑
Single-Hop Retrieval	Random	-	-	-	-	-	13.29	-	37.51
	BM25 (Amati, 2009)	-	-	-	-	-	70.84	-	62.42
	DPR (Karpukhin et al., 2020)	-	-	-	-	-	88.04	-	67.39
Multi-Hop Retrieval	GRR (Asai et al., 2020)	24.92	25.19	5.86	71.45	**99.39**	60.13	25.05	55.20
	SSG (Thorne et al., 2021)	34.72	35.12	6.43	75.81	78.23	77.94	53.85	63.04
	MDR (Xiong et al., 2021)	25.46	25.46	5.82	84.72	97.96	79.97	62.83	51.26
LLMs' Retrieval	GPT-3 (Brown et al., 2020)	0.07	17.14	8.04	12.75	23.97	10.62	0.07	37.12
	GLM (Zeng et al., 2022)	0.68	4.76	7.09	11.16	21.26	8.67	0.68	38.02
	Instruct-GPT (Ouyang et al., 2022)	35.91	54.77	**1.18**	71.92	67.47	81.79	40.78	56.49
Graph-Hop Retrieval	*only w/* Graph-Hop's Forward	27.71	28.86	6.72	88.78	87.2	92.81	67.12	69.71
	only w/ Graph-Hop's Backward	56.57	57.86	4.64	85.67	85.12	88.55	64.06	67.60
	BGR	**56.71**	**58.48**	4.70	**91.81**	90.785	**95.23**	**68.82**	**70.18**
Human Bound		92.15	93.15	0.18	98.13	98.73	97.54	96.41	95.13

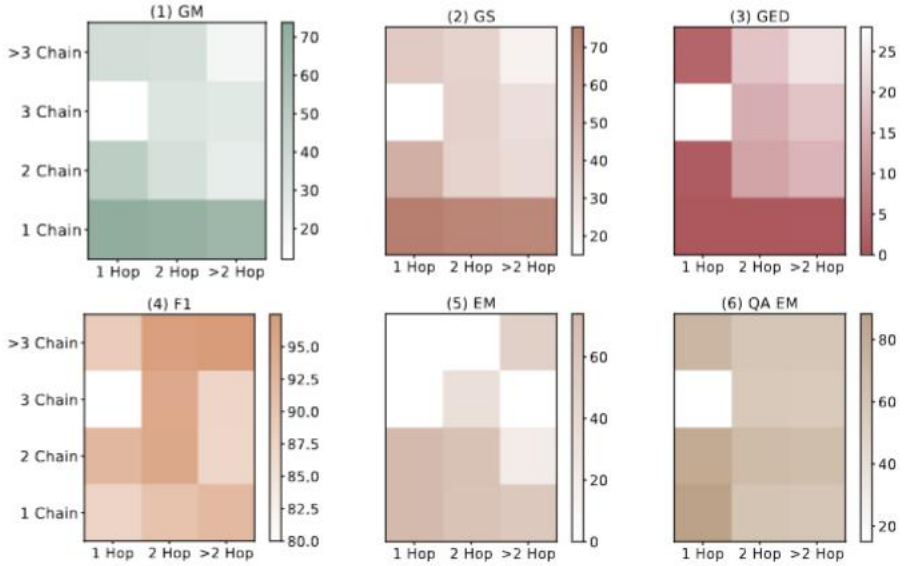
Conducted comprehensive tests to evaluate the performance of our BGR model against traditional retrieval models.

Key metrics used included F1 score, Precision, Recall, and Graph Matching (GM).

Futher Analysis



The retrieval performance for different question and graph structure types.



Experimental results of BGR at different hops and different chain numbers.

TOWARDS GRAPH-HOP RETRIEVAL AND REASONING IN COMPLEX QUESTION ANSWERING OVER TEXTUAL DATABASE

Thanks !

Github: https://github.com/zhu-minjun/Graphhop