HOLMES: Hyper-Relational Knowledge Graphs for Multi-hop Question Answering using LLMs

Pranoy Panda[1] Ankush Agarwal[1] Chaitanya Devaguptapu[1]
Manohar Kaul[1] Prathosh AP[1,2]
[1]Fujitsu Research of India [2]Indian Institute of Science, Bengaluru

Association for Computational Linguistics
ACL 2024

Motivation

LLMs alone (with RAG) can answer simple questions!

"How many board meetings were held in the last twelve months?"

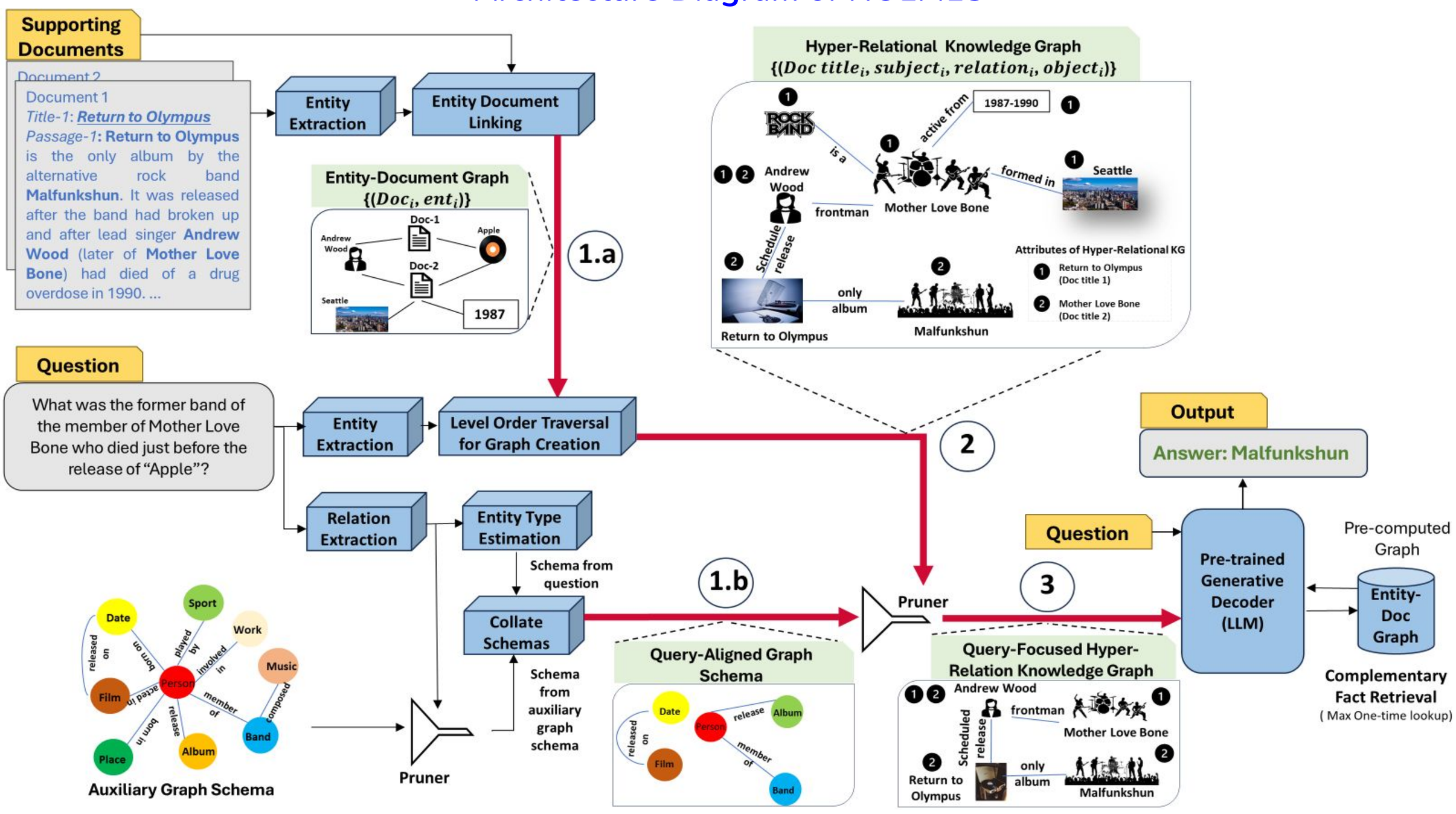
However, real-world problems require solving more complex questions.

"For the board meeting with the most divided votes in the last twelve months, what was the agenda, who voted against it, and by what margin did it pass or fail?"

Contributions

- A new multi-hop QA approach that transforms unstructured text into a hyper-relational KG using a query-derived schema, serving as an input to the LLM.
- A significant improvement over the SoTA multi-hop QA method StructQA (Li and Du, EMNLP 2023) and standard RAG
- HOLMES uses 67% fewer tokens than the current SoTA method by retaining only query relevant information

Architecture Diagram of HOLMES



Multi-Hop Question Answering Performance

Datasets	HotpotQA				MuSiQue			
Methods	EM (↑)	F1 (↑)	P (↑)	R (↑)	EM (↑)	F1 (↑)	P (↑)	R (↑)
Reader: gpt-4-1106-preview								
Base (w/o supp docs)	0.26	0.45	0.45	0.50	0.09	0.21	0.22	0.21
Base (with supp docs)	0.54	0.74	0.75	0.77	0.39	0.55	0.55	0.56
StructQA (Li and Du, 2023)	0.55	0.77	0.75	**0.80**	0.42	0.56	0.57	0.56
Our Method	**0.66**	**0.78**	**0.80**	0.79	**0.48**	**0.58**	**0.59**	**0.59**
Reader: gpt-3.5-turbo-1106								
Base (w/o supp docs)	0.23	0.37	0.38	0.40	0.06	0.15	0.17	0.15
Base (with supp docs)	0.47	0.65	0.66	0.68	0.24	0.36	0.36	**0.37**
StructQA (Li and Du, 2023)	0.48	0.64	0.62	0.67	0.23	0.37	0.37	**0.37**
Our Method	**0.57**	**0.69**	**0.69**	**0.70**	**0.29**	**0.38**	**0.39**	**0.37**

Datasets	HotpotQA		
Methods	EM (↑)	F1 (↑)	SA-EM (↑)
Reader: Gemini-Pro			
Base (w/ supp docs)	0.48	0.66	0.48
StructQA	0.49	0.66	0.52
Our Method	**0.58**	**0.67**	**0.66**

Datasets: HotpotQA, MuSiQue
Baselines: StructQA [EMNLP 2023], Vanilla RAG
LLMs: GPT-4, GPT-3.5, Gemini-pro
Metrics: Exact Match, F1 score, Precision, Recall. We perform upto *20% better on HotpotQA and 14.3% better on MuSiQue* with GPT-4.

Analysis

Datasets	HotpotQA (100 samples)			
Methods	EM (↑)	F1 (↑)	P (↑)	R (↑)
Triple Extractor: gpt-4-1106-preview				
StructQA	0.56	0.76	0.79	0.77
Our Method	**0.68**	**0.79**	**0.82**	**0.80**
Triple Extractor: gpt-3.5-turbo-1106				
StructQA	0.47	0.71	0.72	0.76
Our Method	**0.61**	**0.77**	**0.78**	**0.78**

Impact of Triple Extractor LLM