Best of Both Worlds: A Hybrid Approach for Multi-Hop Explanation with Declarative Facts

Shane Storks, Qiaozi Gao, Aishwarya Reganti, Govind Thattai
Amazon Alexa AI

Introduction and Problem Statement

Efficient, natural human communication relies on implicit shared knowledge and underlying reasoning processes. Despite rapid progress in language-enabled AI agents for tasks like question answering, state-of-the-art systems struggle to explain decisions in natural language. In this work, we present novel multi-hop explanation approaches which integrate efficient syntactic retrieval methods with flexible semantic modeling methods.

We explore the problem of generating multi-hop explanations to support the answer to a natural language question, where the explanation chain is generated from an unstructured corpus of declarative facts. Unstructured natural language corpora are suitable knowledge resources for human-AI interaction, as humans can easily support reasoning by providing their own commonsense knowledge in short, natural language statements. This carefully restricted problem of explanation generation consists of two key challenges:

1. **Retrieval** of candidate supporting evidence from the corpus
2. **Multi-hop reasoning** to connect pieces of evidence into a valid explanation to justify the answer

Datasets

Question Answering via Sentence Composition (QASC) consists of about 10,000 multiple-choice science questions, where each question requires composing 2 facts from a corpus of about 17 million declarative facts to connect the question and its answer [1]:

Q: Differential heating of air can be harnessed for what?

A: electricity production

Explanation:
1. Differential heating of air produces wind.
2. Wind is used for producing electricity.

Figure 1: Example of a question (Q), answer (A), and 2-hop explanation from QASC [1].

QASC includes a gold, human-curated explanation from the corpus for each question-answer pair. We tackle the difficult task of **generating valid explanations from the QASC Corpus**. We compare systems' success on this task by their **gold retrieval rate**: the percentage of question-answer pairs for which the gold explanation was successfully reproduced.

Explainable QASC [2] generates 10 additional explanations for each question-answer pair (hand-labeled as valid or invalid), providing a baseline for the task.

Syntactic Explanation

We indexed the corpus into ElasticSearch, a fast syntactic search engine based on keyword overlap. Following [2], we used a simple **syntactic explanation pipeline** to generate 2-hop explanations for QASC:

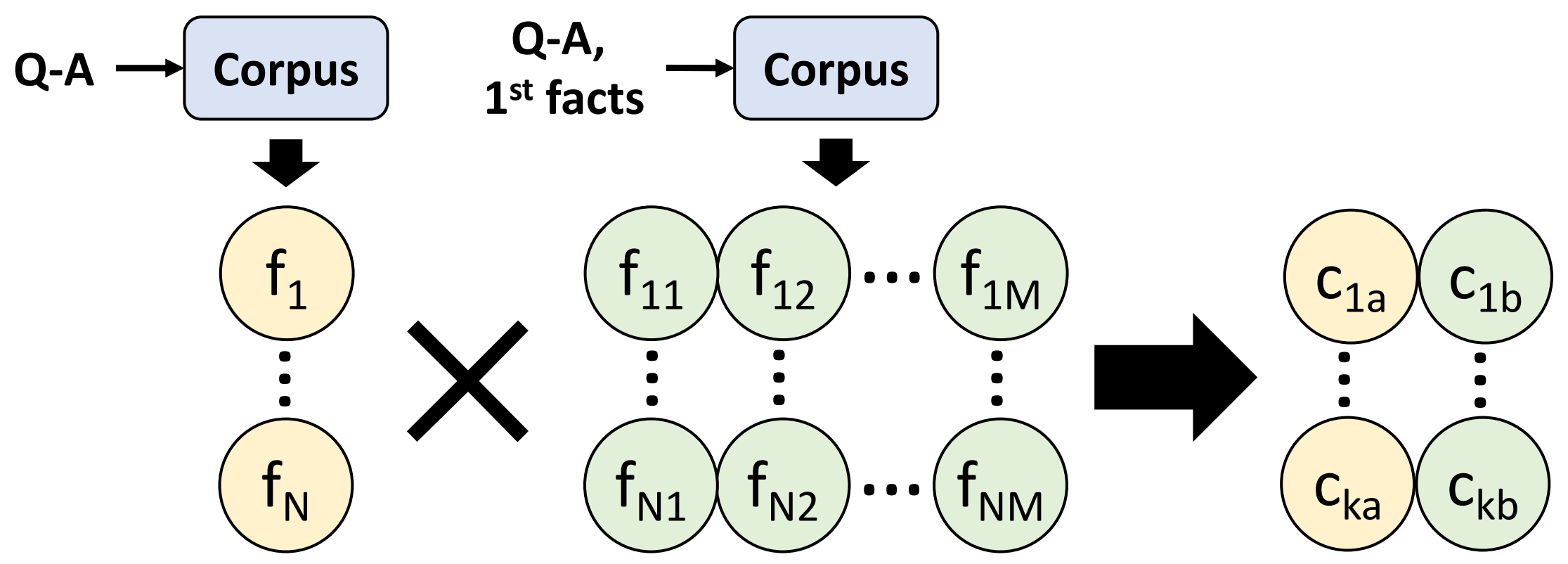

Figure 2: We use the question-answer (Q-A) pair to query the corpus for N facts, each of which is used to query for M second facts. The top-scored K fact pairs are returned.

By increasing the hyperparameters (K, N, M) from their original values $(20, 4, 10)$, we expand and diversify the search, improving the gold retrieval rate:

N	M	K	Gold Retrieval Rate (%)
20	4	10	31.1
20	4	200	37.0
200	200	200	**46.5**

Table 1: Gold explanation chain retrieval rates for syntactic multi-hop explanation on QASC validation set.

Semantic Explanation

We then indexed the corpus into a dense passage retrieval (DPR) index [3], which we trained using question-answer pairs and gold explanation chains from QASC. To reduce error accumulation when generating a multi-hop explanation with this approach, we additionally train a lightweight, feedforward fact **re-encoder**. Given the learned embeddings of a question-answer pair and the first fact from its gold explanation, the re-encoder is trained to generate a new fact embedding similar to that of the gold second fact. At inference time, we use the following **semantic explanation pipeline** to generate 2-hop explanations for QASC:

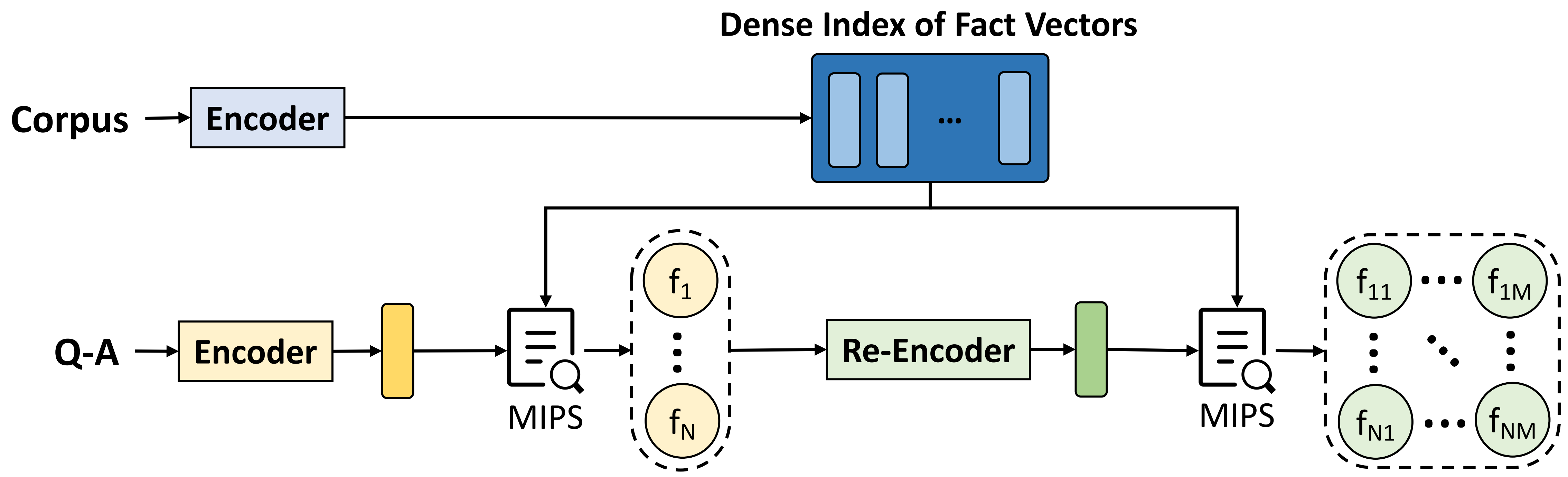

Figure 3: Semantic explanation pipeline. Facts are encoded using a fact encoder (in blue) and stored in a dense index, while the question-answer pair is encoded by a query encoder (in yellow). Maximum inner product search (MIPS) is used to query the index for N candidate first facts, which are each re-encoded (in green), then used to query the index again for M candidate second facts. All candidate first and second facts are paired, and the top-scored K chains are returned as explanations.

In **Table 2**, we compare the gold retrieval rate of our semantic pipeline with the syntactic pipeline, our expanded syntactic pipeline, and a hybrid pipeline using facts retrieved by both the expanded syntactic and semantic pipelines. Among the top 200 proposed explanations by each approach, our hybrid approach improves the gold retrieval rate by up to 12.9%.

Approach	Gold Retrieval Rate (%)	
	Validation	*Test*
syntactic	37.0	40.2
syntactic (exp.)	46.5	49.3
semantic	10.8	13.9
syntactic (exp.) + semantic	**49.9**	**51.1**

Table 2: Gold retrieval rates (top $K = 200$ candidates) for combinations of multi-hop retrieval approaches on QASC.

Re-Ranking Explanations

As a set of $K = 200$ explanations for a single question-answer pair may become unpractical for human use, we lastly apply an **explanation re-ranker** to filter the proposed explanations to only the best $K = 10$ candidates. To implement the re-ranker, we fine-tune pretrained language models [4, 5] to classify whether an explanation chain is valid for a question-answer pair. As shown in **Table 3**, when compared to the syntactic approach used in [2], our hybrid approach improves the gold retrieval rate by up to 7% after re-ranking.

Retrieval Approach	Re-Ranker	Gold RR (%)	
		Val.	*Test*
syntactic [2]	–	31.1	34.1
syntactic (exp.)	BERT	36.3	34.0
syntactic (exp.) + semantic	BERT	36.4	34.1
syntactic (exp.)	RoBERTa	37.9	36.2
syntactic (exp.) + semantic	RoBERTa	**38.1**	**36.4**

Table 3: Gold retrieval rates (top $K = 10$ candidates) for combinations of multi-hop retrieval approaches on QASC, re-ranked by fine-tuned language models [4, 5].

Conclusion

In this work, by utilizing a small amount of ground truth supervision, we explored approaches to improve the generation of multi-hop explanations from a corpus of declarative facts. We showed that both fast, syntactic methods and slow, semantic methods are useful for gathering relevant evidence for explanation.

References

[1] T. Khot, P. Clark, M. Guerquin, P. Jansen, and A. Sabharwal. QASC: A dataset for question answering via sentence composition. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, 2020.

[2] H. Jhamtani and P. Clark. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019.