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

Abstract

Language-enabled AI systems can answer complex, multi-hop questions to high accuracy, but supporting answers with evidence is a more challenging task which is important for the transparency and trustworthiness to users. Prior work in this area typically makes a trade-off between efficiency and accuracy; state-of-the-art deep neural network systems are too cumbersome to be useful in large-scale applications, while the fastest systems lack reliability. In this work, we integrate fast syntactic methods with powerful semantic methods for multi-hop explanation generation based on declarative facts. Our best system, which learns a lightweight operation to simulate multi-hop reasoning over pieces of evidence and fine-tunes language models to re-rank generated explanation chains, outperforms a purely syntactic baseline from prior work by up to 7% in gold explanation retrieval rate.

Introduction

Efficient and 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 and more, state-of-the-art systems still struggle to explain their decisions in natural language. To improve their interpretability and robustness, a number of multi-hop explanation generation and identification benchmarks based on large, unstructured corpora of facts have been created (Mihaylov et al. 2018; Khot et al. 2020; Jhamtani and Clark 2020). However, when generating explanation chains, powerful deep neural networks can be too cumbersome to be useful in large-scale applications, while the fastest systems lack reliability, as they rely on syntactic features and ignore semantic relations between concepts (Banerjee and Baral 2020; Jhamtani and Clark 2020). In this work, we present novel approaches to integrate efficient syntactic retrieval methods with flexible semantic modeling methods for multi-hop explanation. Our methods simulate a multi-hop reasoning process from the retrieval and synthesis of evidence to re-ranking candidate explanations.

Related Work

Recent work has focused on different aspects of multi-hop reasoning for question answering and related natural language understanding tasks. One line of work has incorporated highly structured knowledge graphs into language understanding by combining graphical methods with language models (Lin et al. 2019; Ji et al. 2020; Yasunaga et al. 2021), augmenting language model inputs with relational knowledge (Zhang et al. 2019; Chen et al. 2020; Xu et al. 2021), and applying language models to relational knowledge to infer multi-hop reasoning paths through knowledge graphs (Wang et al. 2020). Others have further explored training language models with semi-structured relational knowledge (Sap et al. 2019; Bosselut et al. 2019; Mostafazadeh et al. 2020; Hwang et al. 2021), i.e., where nodes are natural language sentences rather than canonicalized concepts, to later use for generating multi-hop explanations in natural language (Shwartz et al. 2020; Bosselut et al. 2021).

For generating multi-hop explanations from entirely unstructured corpora, other work has explored using multi-step syntactic information retrieval methods (Jhamtani and Clark 2020), and modeling such corpora as knowledge graphs with relations induced by shared mentions of concepts between documents (Dhingra et al. 2020; Lin et al. 2021). While the former approach lacks the ability to capture semantic relationships between evidence sentences, the latter demands high time and space complexity both in generating a graph from corpora of millions of facts, and in everyday uses of adding or removing facts from the corpus. More recent work has used pre-trained word embeddings to add some lightweight semantic representation to syntactic evidence retrieval (Yadav, Bethard, and Surdeanu 2021). Unlike these approaches, we present a flexible and relatively lightweight pipeline to apply both syntactic and learned, contextualized semantic approaches in multi-hop explanation generation, including evidence retrieval, multi-hop reasoning over evidence, and re-ranking candidate explanations.

Problem Statement

In the research community, two types of explanation have been studied: introspective explanation and justification explanation (Biran and Cotton 2017). The former explicates how a decision is made, and the latter gathers evidence to support a decision. In this study, we focus on the task of justification explanation. Specifically, we explore the problem of generating multi-hop explanations to support the answer to a natural language question, where the explanation

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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. First, we must solve the *retrieval* task to gather candidate supporting evidence from the corpus. Second, we need to invoke a *multi-hop reasoning* process to connect pieces of evidence to form the most valid explanation to justify the answer to the question.

Datasets

To explore this problem, we consider two datasets. First, the Question Answering via Sentence Composition (QASC) dataset provides a about 10,000 multiple-choice science questions (Khot et al. 2020). QASC is a challenging problem, as each question requires composing two facts from a corpus of about 17 million declarative facts to connect the question and its correct answer. For example, given the question "*Differential heating of air* can be harnessed for what?" and correct answer "*electricity production*," the answer can be explained by composing the facts "*Differential heating of air* produces wind" and "Wind is used for *producing electricity*," which connect the question and answer. QASC includes a gold, human-curated 2-hop explanation from the corpus for each question-answer pair.

Meanwhile, the Explainable QASC (eQASC) dataset adds 10 automatically generated explanations for each question-answer pair, each of which are labeled by annotators as valid or invalid (Jhamtani and Clark 2020). While the state-of-the-art accuracy on QASC has reached up to 90%,[1] only 76% of questions have any valid explanation chains in eQASC. This indicates that *explaining* the answers to questions in QASC is a more challenging problem than answering them. This motivates us to further explore the problem of generating multi-hop explanations for QASC.

Methods

In our experiments toward multi-hop explanation generation, we consider syntactic and semantic multi-hop retrieval methods, then explore ways to re-rank retrieved explanations to reduce the pool of candidates.

Syntactic Methods

Syntactic information retrieval methods enable quick searching of millions of documents. eQASC was originally generated using ElasticSearch [2], a fast but primarily syntactic search engine based on keyword overlap. After indexing the QASC corpus facts into an ElasticSearch index, Jhamtani and Clark (2020) used a simple procedure (shown in Figure 1) to generate a 2-hop explanation for each question-answer pair from QASC. First, query the corpus for $N = 20$ candidate first facts. For each candidate first fact, query the corpus for $M = 4$ candidate second facts, where each candidate second fact must contain a word that appears in the question-answer pair, and a word that appears in the first fact. The purpose of this restriction is to force the resulting chain of facts to connect concepts in the question and answer through intermediate concepts. Lastly, from the set product of all candidate first facts and all candidate second facts, select up to $K = 10$ candidate explanation chains, ranked by the sum of retrieval scores from the ElasticSearch engine.

[1] See https://allenai.org/data/qasc.

[2] https://www.elastic.co/

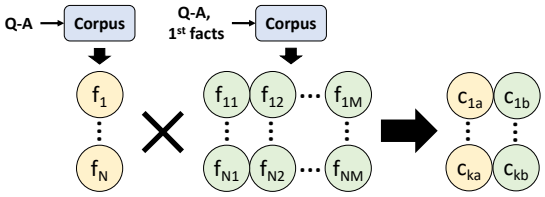

Figure 1: Syntactic pipeline used to generate multi-hop explanations in eQASC. First, the question-answer (Q-A) pair is used to query the ElasticSearch index for N candidate first facts, each of which is used to query it for M candidate second facts. All candidate first and second facts are paired, and the top-scored K chains are returned as explanations.

Expanding syntactic retrieval. This is a simple, fast approach to generate a large number of candidate explanation chains. To improve the likelihood that a valid explanation is generated, we can expand and diversify the search results by increasing N, M, and K. Specifically, we increase N, M, and K each to 200.

Semantic Methods

Alternatively, semantic information retrieval methods can enable stronger meaning representation than syntactic methods with a trade-off of search speed. Typical approaches generate a semantic vector embedding for all documents in a corpus. They then generate a comparable embedding of the query, and use simple vector similarity measures, e.g., inner product, to rank documents.

Dense passage retrieval. Dense passage retrieval (DPR) is a recent approach to semantic information retrieval which learns dual encoders for queries and documents (Karpukhin et al. 2020). They are trained such that the query and document encoders generate similar embeddings for semantically similar queries and documents. Similarity is measured by inner product of vectors, and is maximized for matching queries and documents, but minimized for irrelevant queries and documents. We can then use the document encoder to index the facts in a corpus, and efficiently query the index using the embedding from the query encoder.

For each question-answer pair in QASC and the two facts in its gold explanation chain, we can train a dense passage retriever to generate similar embeddings for the question-answer pair (query) and these facts (documents), then encode all facts in the QASC corpus for search purposes.

Multi-hop reasoning. To generate a multi-hop explanation using this approach, we need to facilitate reasoning over the facts and queries. Given a question-answer pair from QASC, we first query the DPR index for $N = 5$ facts. To reduce error accumulation in generating the chain of facts,

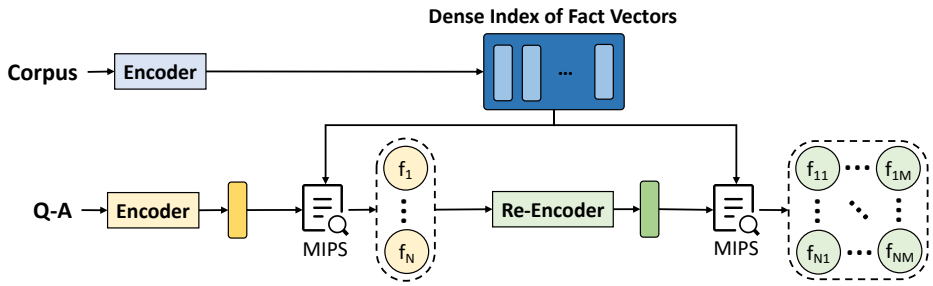

Figure 2: Proposed semantic explanation pipeline. Facts are encoded using a fact encoder (in blue) and stored in a dense index, while the question-answer pair is encoded by a query encoder (in yellow). Maximum inner product search (MIPS) is used to query the index for N candidate first facts, which are 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.

we then *re-encode* each fact into a new query embedding incorporating the candidate fact and the original query.

The re-encoder is a lightweight feedforward network inspired by the fact-translating function proposed in Lin et al. (2021). Given the embeddings q_{QA} and d_1 for the question-answer pair query and first fact document respectively, we use the gold explanation chains from QASC to learn the re-encoder $g(q_{QA}, d_1)$. Specifically, if d_1 is the embedding of the first fact in the gold explanation chain, we maximize the inner product between the re-encoded output q_r and the document embedding d_2 for the second gold fact. Next, we query the DPR index again using q_r to obtain $M = 2$ candidate second facts. To reduce noise, we filter out any facts that mention no concepts in either the question or answer. Lastly, from the set product of all candidate first facts and all candidate second facts, select up to $K = 10$ candidate explanation chains, ranked by the sum of retrieval scores, i.e., inner products when querying the DPR index. Our semantic explanation pipeline is shown in full in Figure 2. It is worth noting that our lightweight re-encoder operation can extend to any number of hops.

Re-Ranking Candidate Explanations

Both of our syntactic and semantic multi-hop retrieval systems can quickly propose candidate explanation chains for questions in QASC. However, both approaches over-generate candidates, and the high number of candidates (i.e., up to 200) limits the practical usefulness of the systems to an end user. As such, we lastly propose a re-ranker for candidates based on large-scale, pre-trained language models (Devlin et al. 2019; Liu et al. 2019). Specifically, we use the gold explanation chains from QASC to fine-tune a language model to the classification task of whether or not a candidate explanation is valid for a question-answer pair. We then re-rank a pool of candidates based on the system's estimated likelihood that each explanation chain is valid, and keep only the top $K = 10$ candidates for a direct comparison to the syntactic approach used to generate eQASC.

Experimental Results

We next apply these approaches to the task of selecting 2-hop reasoning chains for question-answer pairs in QASC, and directly compare our results to the original procedure to generate eQASC. We compare systems by their individual *gold retrieval rate* on the validation set for QASC, i.e., the percentage of question-answer pairs for which the gold explanation chain from QASC was successfully reproduced.[3] This serves as an indicator of the quality of generated explanations, as it suggests that generated explanations tend to look more like those curated from the corpus by humans.

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 retrieval with ElasticSearch on QASC validation set. The first row indicates the original search hyperparameters used to generate eQASC, while the last two rows increase hyperparameters to expand and diversify the search.

Expanded Syntactic Explanation

As mentioned earlier, we first expanded the ElasticSearch-based approach used to generate eQASC by increasing the search hyperparameters N, M, and K each to 200. Selected results from adjusting these hyperparameters are listed in Table 1. By only increasing K (i.e., the number of candidate explanation chains considered) to 200, the retrieval rate increases from 31.1% to 37.0%. When increasing N and M (i.e., the number of candidate first and second facts considered) also to 200, the retrieval rate further increases to 46.5%, a net 15.4% gain.

Syntactic-Semantic Multi-Hop Explanation

Next, we incorporate our semantic multi-hop retrieval process powered by DPR.

Training details. The dual encoders for DPR are learned starting from pre-trained BERT-base (Liu et al. 2019). The best encoders are selected based on the mean squared error between embeddings for matching question-answer pairs and facts on the QASC validation set. The batch size is fixed at 16, while learning rate and number of training epochs are

[3] As the ordering of facts in QASC explanation chains does not typically matter, the gold retrieval rate counts both the forward and reverse forms of gold explanation chains.

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 explanation chain retrieval rates (top $K = 200$ candidates) for combinations of multi-hop retrieval approaches on QASC. Syntactic refers to the second result from Table 1, while syntactic (exp.) refers to the expanded third result. Semantic refers to the previously introduced DPR-based multi-hop retrieval approach.

selected based on a grid search. For the re-encoder, the training batch size, learning rate, and number of epochs are similarly selected based on a grid search, minimizing the mean squared error between the output re-encoded queries and target fact embeddings on the validation set.

Results. Table 2 compares the gold retrieval rate of various combinations of the syntactic and semantic approaches for multi-hop retrieval on the QASC validation and testing sets.[4] Our results show that while using only the semantic candidate explanation chains leads to a 13.9% gold retrieval rate at best, combining the expanded syntactic and semantic candidates gives us the best result of up to 51.1% gold retrieval rate, outperforming the case where only the expanded syntactic candidates are considered. Thus, the semantic approach finds some of the missing gold explanations that the syntactic approach misses, suggesting that both syntactic and semantic approaches are needed for generating the best-quality explanations on QASC questions.

LM Re-Ranking

While our results improve the gold retrieval rate by a wide margin, recall that our multi-hop retrieval approaches for QASC increase the number of candidate explanation chains K to 200. Such a large set of candidates is not useful in practice, as a human user would have to sort through a cumbersome number of explanations in order to judge the machine's understanding of the question and answer. As such, we lastly present our experiments on *re-ranking* candidate explanation chains, which enables us to truncate our results to $K = 10$ top candidate explanation chains without massive performance drops, and consequently compare our approach directly to the original approach used to generate eQASC.

Training details. Using the re-ranking approach described earlier, we fine-tune the BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) pre-trained language models.[5] Models are trained with a 1:2 ratio of gold and invalid explanation chains, with 3 unique invalid explanation chains randomly sampled from ElasticSearch results per gold explanation chain (forward and reverse forms). Models are selected based on instances achieving the highest top-1 gold retrieval rate, i.e., proportion of question-answer pairs where the gold explanation chain is ranked highest, on the QASC validation set similarly redistributed in this way.

Retrieval Approach	Re-Ranker	Gold RR (%)	
		Val.	*Test*
syntactic	–	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 explanation chain retrieval rates (RR; top $K = 10$ candidates) for combinations of multi-hop retrieval approaches on QASC, re-ranked by fine-tuned language models. Syntactic refers to the original approach used to generate eQASC, while syntactic (exp.) refers to the expanded third result from Table 1. Semantic refers to our proposed DPR-based multi-hop retrieval approach.

Results discussion. Table 3 compares the final gold retrieval rates for the top $K = 10$ re-ranked candidates from various approaches. While the original syntactic approach for generating eQASC achieves a respective 31.1% and 34.1% gold retrieval rate on the validation and testing sets, our expanded syntactic approach achieves up to 37.9% and 36.2% gold retrieval rate with RoBERTa. Again with RoBERTa, our syntactic-semantic multi-hop retrieval achieves the best results of 38.1% and 36.4% on the validation and testing sets, respectively, exceeding the baselines. While our best approach achieves up to a net 7.0% gain over the baseline syntactic approach used to generate eQASC, much of our performance gain was lost after re-ranking. This is possibly because the negative examples proposed by DPR were not incorporated in re-ranker training, which may cause generalization issues. Future work may further explore the re-ranking problem to better preserve the retrieval rates from the larger candidate pools explored earlier.

Conclusion

In this work, by utilizing a small amount of ground truth supervision, we explored several approaches to improve the generation of multi-hop explanations from a corpus of declarative facts. Our results show that both fast, syntactic methods and slow, semantic methods are useful for gathering relevant evidence for explanation. To facilitate multi-hop reasoning from one piece of evidence to the next, we had some success in using a lightweight feedforward re-encoder, as opposed to state-of-the-art graph-based approaches that consume too much time and memory for practical online use. As many of our results over-generate candidate multi-hop explanation chains, we lastly explored using pre-trained language models to re-rank and filter candidates to only the best ones. However, our results show that this is a significant challenge, and future work may further explore this problem.

[4]When combining the syntactic and semantic approaches, we replace up to the lowest-ranked 25% of syntactic candidate explanation chains with the top semantic candidate explanation chains.

[5]For both models, we use the "base" form which has 12 hidden layers, a hidden dimension of 768, and 12 attention heads.

References

Banerjee, P.; and Baral, C. 2020. Knowledge fusion and semantic knowledge ranking for open domain question answering. *arXiv preprint arXiv:2004.03101*.

Biran, O.; and Cotton, C. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, 8–13.

Bosselut, A.; Bras, R. L.; ; and Choi, Y. 2021. Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-shot Commonsense Question Answering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4762–4779. Florence, Italy: Association for Computational Linguistics.

Chen, Q.; Ji, F.; Chen, H.; and Zhang, Y. 2020. Improving Commonsense Question Answering by Graph-based Iterative Retrieval over Multiple Knowledge Sources. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2583–2594. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dhingra, B.; Zaheer, M.; Balachandran, V.; Neubig, G.; Salakhutdinov, R.; and Cohen, W. W. 2020. Differentiable Reasoning over a Virtual Knowledge Base. In *International Conference on Learning Representations*.

Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs.

Jhamtani, H.; and Clark, P. 2020. 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 (EMNLP)*, 137–150. Online: Association for Computational Linguistics.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; Zhu, X.; and Huang, M. 2020. Language Generation with Multi-Hop Reasoning on Commonsense Knowledge Graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 725–736. Online: Association for Computational Linguistics.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Online: Association for Computational Linguistics.

Khot, T.; Clark, P.; Guerquin, M.; Jansen, P.; and Sabharwal, A. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8082–8090.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2829–2839. Hong Kong, China: Association for Computational Linguistics.

Lin, B. Y.; Sun, H.; Dhingra, B.; Zaheer, M.; Ren, X.; and Cohen, W. 2021. Differentiable Open-Ended Commonsense Reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4611–4625. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Mostafazadeh, N.; Kalyanpur, A.; Moon, L.; Buchanan, D.; Berkowitz, L.; Biran, O.; and Chu-Carroll, J. 2020. GLUCOSE: GeneraLized and COntextualized Story Explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4569–4586. Online: Association for Computational Linguistics.

Sap, M.; LeBras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI*.

Shwartz, V.; West, P.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2020. Unsupervised Commonsense Question Answering with Self-Talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4615–4629. Online: Association for Computational Linguistics.

Wang, P.; Peng, N.; Ilievski, F.; Szekely, P.; and Ren, X. 2020. Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4129–4140. Online: Association for Computational Linguistics.

Xu, Y.; Zhu, C.; Xu, R.; Liu, Y.; Zeng, M.; and Huang, X. 2021. Fusing Context Into Knowledge Graph for Commonsense Question Answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1201–1207. Online: Association for Computational Linguistics.

Yadav, V.; Bethard, S.; and Surdeanu, M. 2021. If You Want to Go Far Go Together: Unsupervised Joint Candidate Evidence Retrieval for Multi-hop Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, 4571–4581. Online: Association for Computational Linguistics.

Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 535–546. Online: Association for Computational Linguistics.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1441–1451. Florence, Italy: Association for Computational Linguistics.