Cornell Bowers CIS
Computer Science

61 ACL 2023

Rationale-based Approaches

Tutorial on Complex Reasoning in Natural Language
ACL 2023

What are rationales?

Definition of rationales

- Rationales are extractive texts that significantly influence what the output would be.
- Rationales were first introduced in Zaidan et al. (2007)

Input What do people do during their time off from work?

(A) take trips, (B) become hysterical

Output (A) take trips

Rationale models may be Supervised / Unsupervised

- Zaidan et al. (2007) supervise models with rationales
- Lei et al. (2016) proposed self-rationalizing models without rationale supervision, making producing rationales possible for every dataset

Rationale models may be Faithful / Unfaithful

- Rationale models are faithful if they predict outputs given only the rationales
- Rationale models need to be faithful to be deemed as an explainable model

Rationale models may extract Tokens / Sentences

- Tokens for short inputs
- Sentences / paragraphs for long inputs
- Complex reasoning tasks often consist of long inputs, i.e., many (and potentially very long) documents

Rationale models may be Single-hop / Structured

- Single-hop rationale models predict sentences in a rationale independently
- Structured rationale models explicitly consider sentence structures

Rationale models are closely related to Retrieval

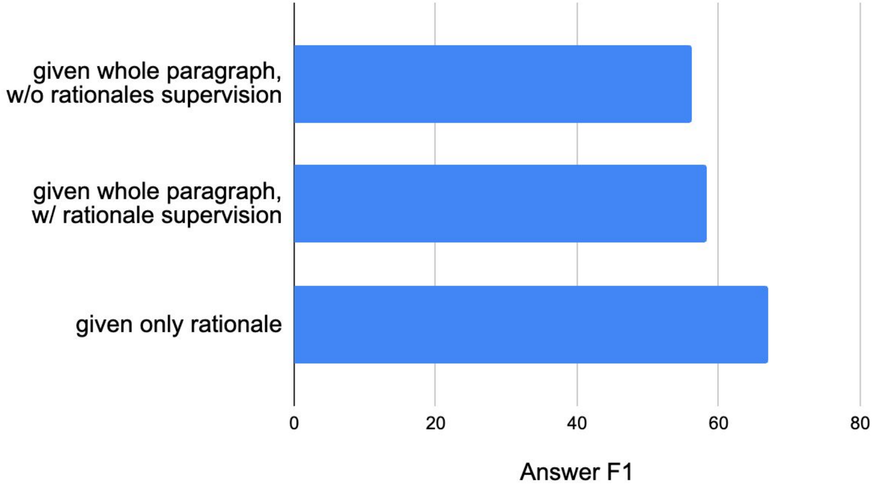
- Documents to be retrieved can be seen as rationales
- Better rationale models can lead to better retrieval models
- More retrieval work is covered in another ACL tutorial: Retrieval-based Language Models and Applications (2pm in the afternoon)

[Lewis et al., 2020]

What are benefits and costs of rationale models?

Benefits of supervised rationale models

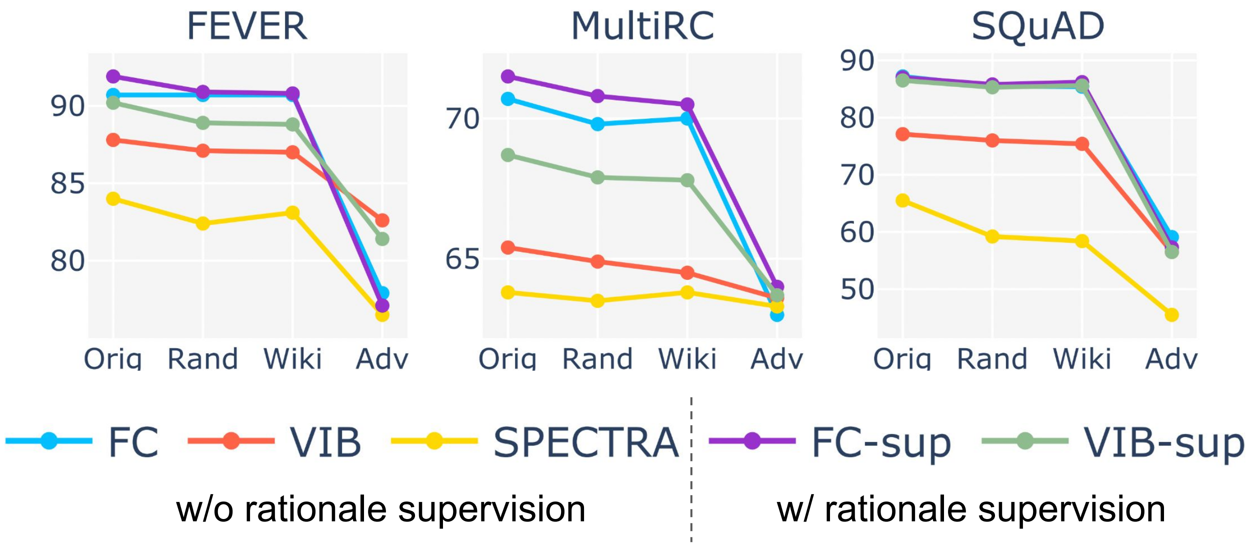
- Rationale models can improve task performance



[Yang et al., 2018]

Benefits of supervised rationale models

- Rationale models are robust to adversarial attacks



[Chen et al., 2022]

Costs of supervising rationales

- Only 29 datasets have annotated rationales [Wiegreffe and Marasović, 2022]
- Rationale annotations are expensive to collect [Geva et al., 2021]
- Rationales can be subjective to annotate [Zhang et al., 2020]

Benefits and costs of structured rationale models

- Necessary to get reasoning correct for problems involve compositional structures
- However, there may be training and / or inference overhead

Benefits of faithful rationale models

- Faithful rationale models allow users to evaluate the trustworthiness of their predictions

Input What do people do during their time off from work?

(A) take trips, (B) become hysterical

Output (A) take trips

[Rajani et al., 2019]

Benefits of faithful rationale models

- Faithful rationale models allow users to debug datasets

Q: Watertown International Airport and Blue Grass Airport, are in which country?

Document A, Blue Grass Airport:
Blue Grass Airport is a public airport in Fayette County, Kentucky, 4 miles west of downtown Lexington.
Document B, Watertown International Airport:
Watertown International Airport is a county owned, public use airport located in Jefferson County, New York, United States.

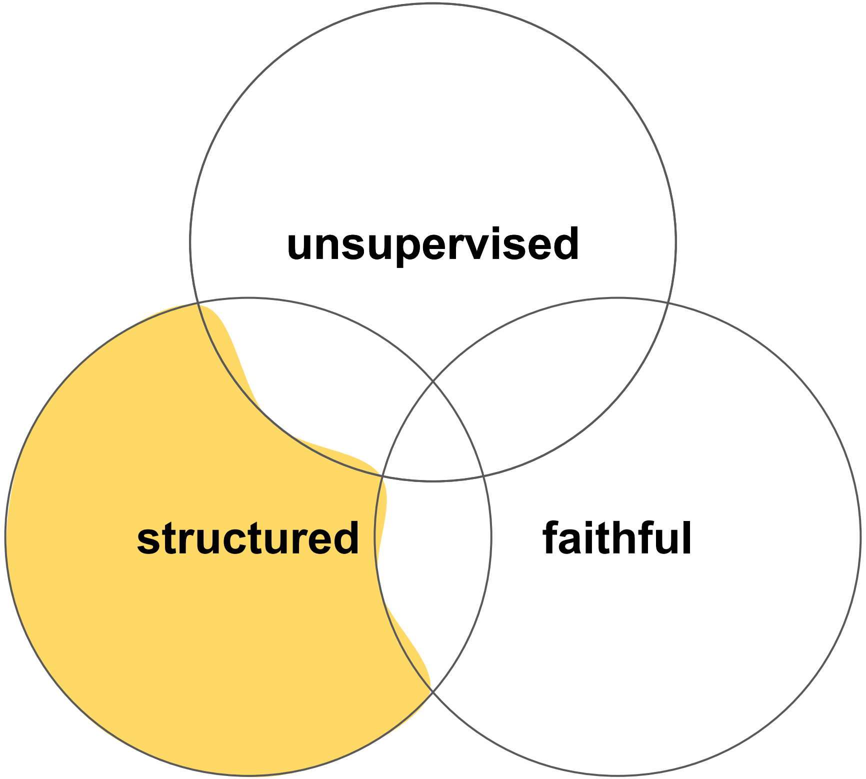
A: United States

[Zhao et al., 2023]

Costs of faithful rationale models

- Potentially more computationally expensive to train
- May not necessarily improve task accuracy

Overview of methods



Transformer models that handle long inputs

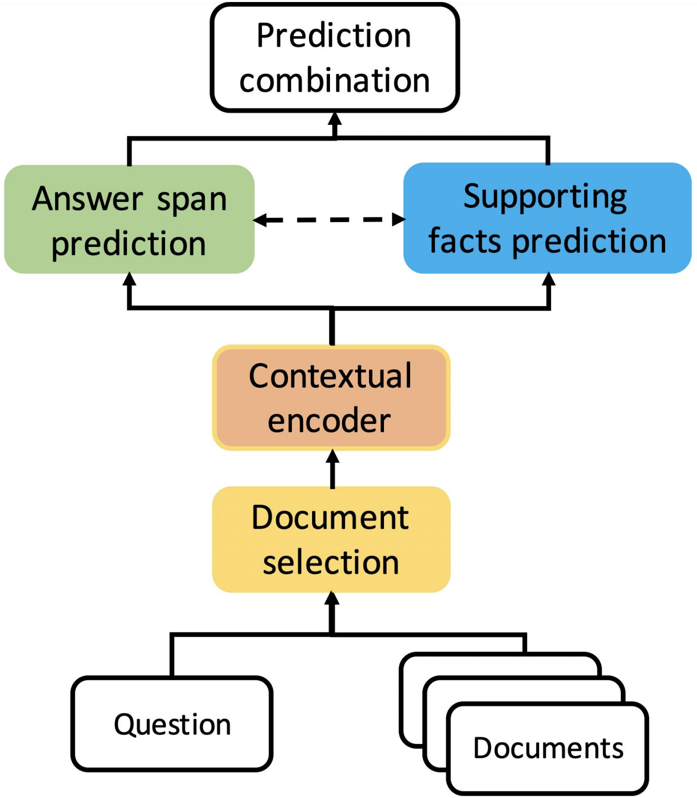
- Feed the entire input into transformer models that handle long-form texts and directly predict a rationale from contextualized embeddings of [SPC] tokens
- Input: {question} [SPC] sent 1 [SPC] … [SPC] sent *n*
- Predict on [SPC] tokens for whether a sentence is included

Input {question} [SPC] sent 1 [SPC] … [SPC] sent *n*

Output 0 / 1? 0 / 1? 0 / 1?

[Beltagy et al., 2020, Zaheer et al., 2020]

Handling long inputs with regular transformers

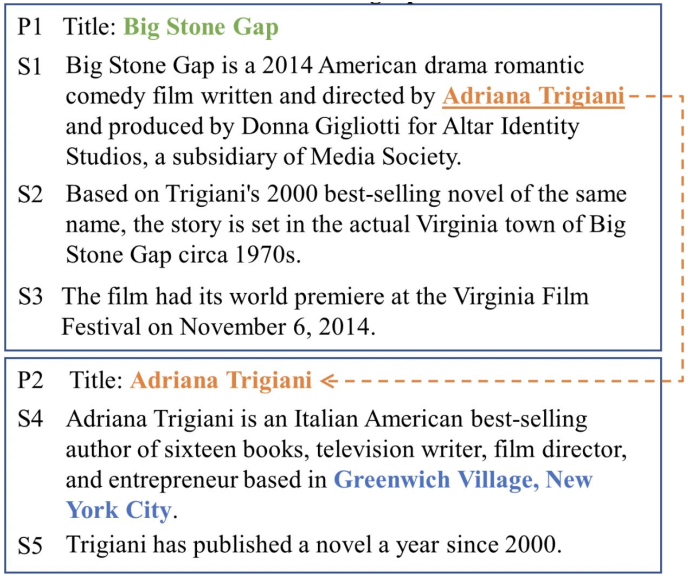


- First, a document selection module filters out answer-unrelated documents
- Then, an answer and explain module, trained with a multi-task loss, jointly predicts an answer and a rationale

[Tu et al., 2019]

Utilizing graph neural networks (GNNs)

- Use graph neural networks to capture the relationship between different hops
- Graphs are often built with entities



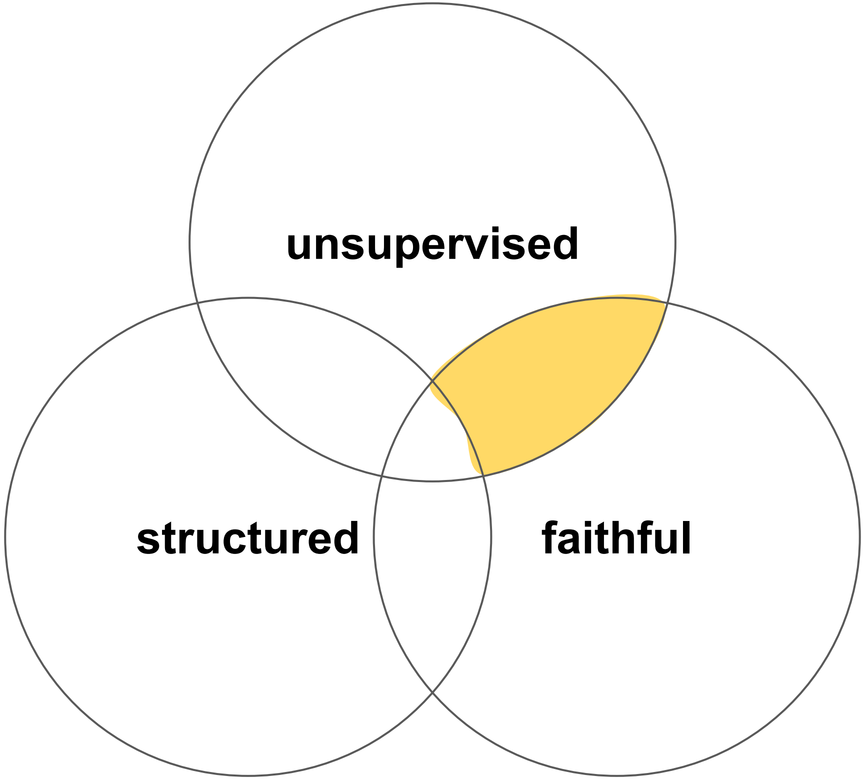
[Thayaparan et al., 2019; Fang et al., 2020; Qiu et al., 2019]

Graph vs. No graph

Model	Answer	Rationale	Joint
w/o Graph	80.58	85.83	71.02
Hier. Graph	**82.22**	**88.58**	**74.37**

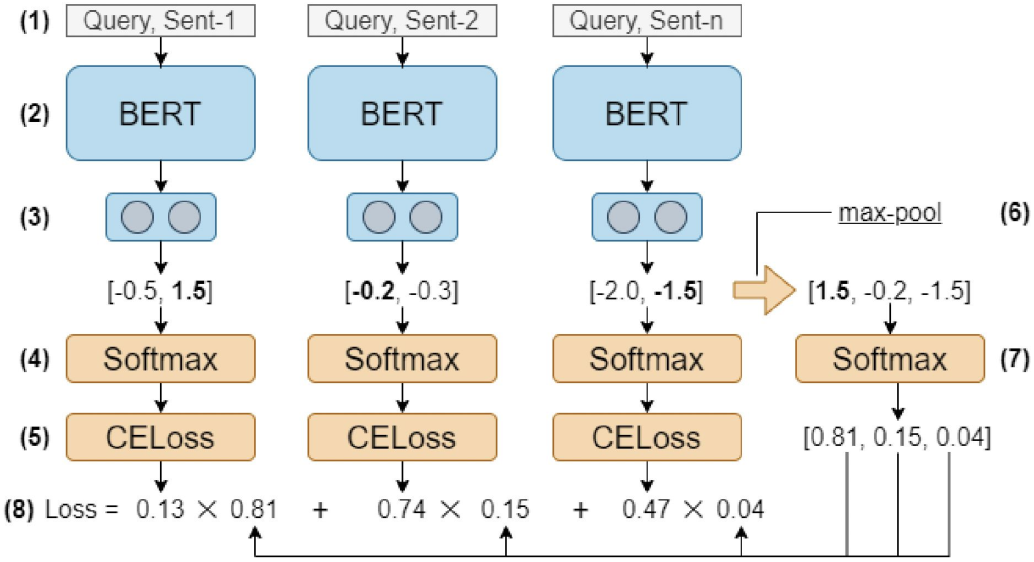
[Fang et al., 2020]

Overview of methods



Prediction confidence

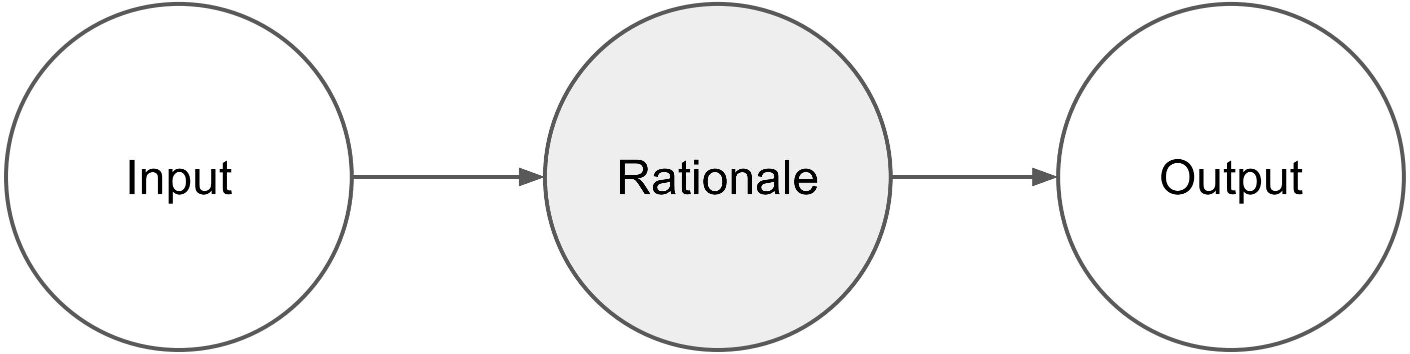
- Treat which part leads to highest prediction confidence as rationale



[Glockner et al., 2020; Atanasova, et al., 2023]

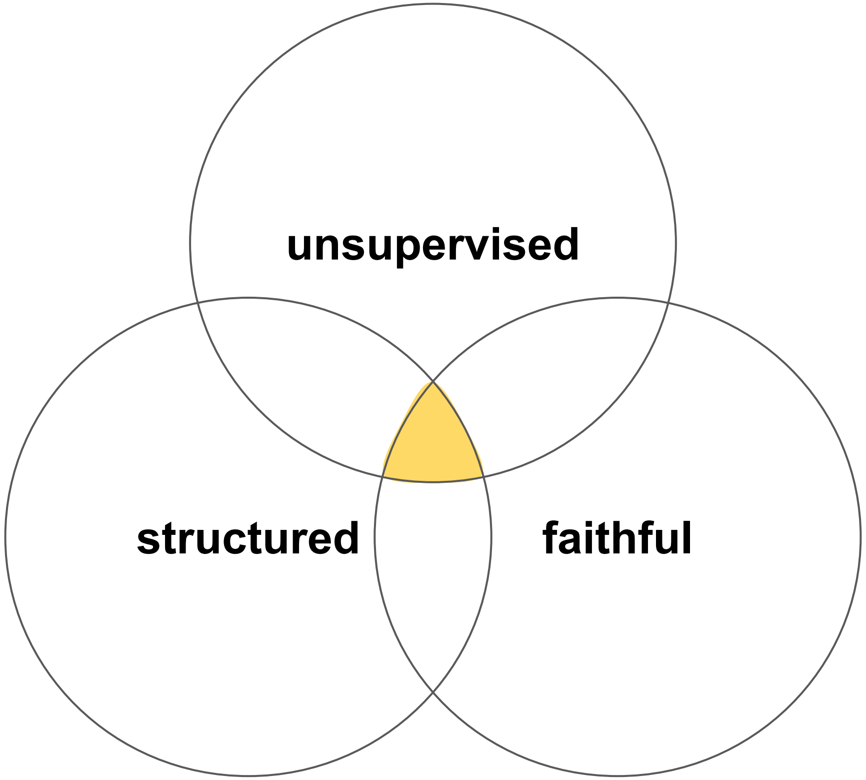
Latent rationales

- Models a single document as a latent variable
- Easy to build: This model is on HuggingFace

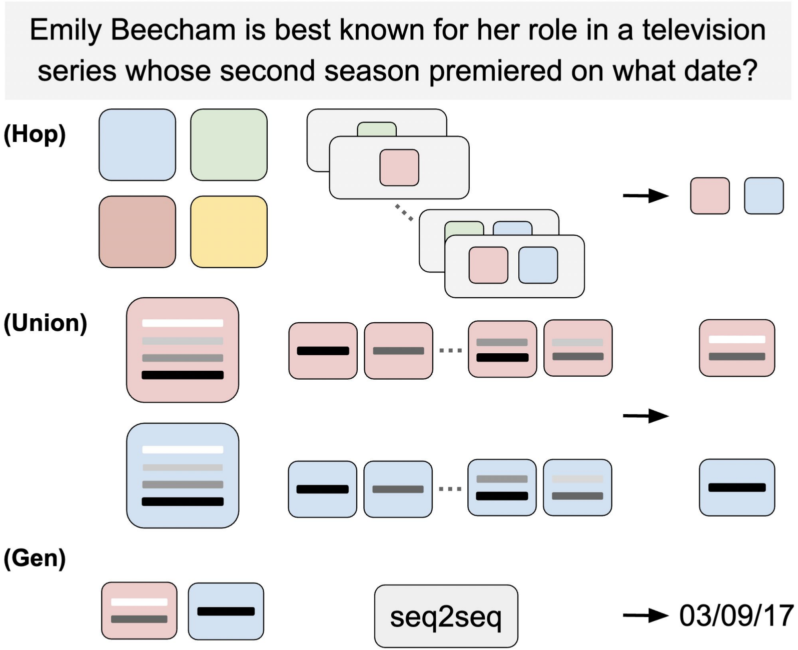


[Lewis et al., 2020]

Overview of methods



Latent set rationales



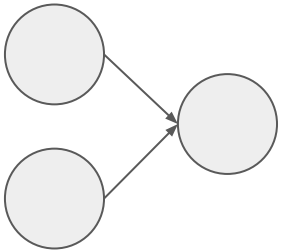
- Explicitly models multi-hop reasoning as set-prediction problems

[Zhao et al., 2023]

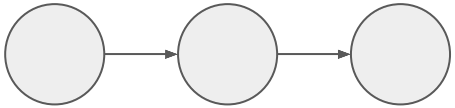
Modeling documents sets vs. single documents

- HUG: models interdependency between documents and sentences
- HUG-ind: models documents and sentences independently

Comparison-based reasoning



Bridge-based reasoning



		Sent F1	Doc F1	Ans F1
Comparison	HUG-Ind	**78.9**	**92.9**	64.8
	HUG	78.1	91.1	**69.7**
Bridge	HUG-Ind	55.2	68.6	71.6
	HUG	**71.0**	**87.3**	**75.7**

Future directions for rationale-based approaches

- How to scale up unsupervised rationale selection remains understudied
- Rationale selection doesn’t automatically solved by larger, better language models, due to long input lengths
- How to explicitly model the structure between sub-rationales?
- How can NLP systems further benefit from rationales?

Conclusion for the tutorial

- Complex reasoning tasks still remains unsolved even with LLMs
- Making reasoning explicit is a promising direction to build NLP systems that generalize and can be trusted by users
- Some of the explicit reasoning systems are easy to implement with open-source tools --- start building today!

Paper list

[Rajani et al., 2019] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

[Trivedi et al., 2022] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. ♫ MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

[Yang et al., 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

[Chen et al., 2022] Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can Rationalization Improve Robustness?. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

Max Glockner, Ivan Habernal, and Iryna Gurevych. 2020. Why do you think that? Exploring Faithful Sentence-Level Rationales Without Supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1080–1095, Online. Association for Computational Linguistics.

Paper list

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

[Thayaparan et al., 2019] Mokanarangan Thayaparan, Marco Valentino, Viktor Schlegel, and André Freitas. 2019. Identifying Supporting Facts for Multi-hop Question Answering with Document Graph Networks. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 42–51, Hong Kong. Association for Computational Linguistics.

[Fang et al., 2020] Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical Graph Network for Multi-hop Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

[Qiu et al., 2019] Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically Fused Graph Network for Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150, Florence, Italy. Association for Computational Linguistics.

[Lewis et al., 2020] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.T., Rocktäschel, T. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, pp.9459-9474.

Paper list

Free-text rationales are excluded

- Free-text rationales, despite being more flexible
    - They might not be faithful to inputs
    - Systems that generate such rationales can be brittle [Camburu et al., (2019)]
    - If you are interested, please refer to Wiegreffe and Marasović [2022]

Rationale models may be Supervised / Unsupervised

- 29 datasets were annotated with rationales

Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Ana Marasović*
Allen Institute for AI
University of Washington
anam@allenai.org