# An incomplete snapshot of multi-hop inference

(@ EMNLP 2021)

TeCho (reading group) Nov. 5

Patrick Xia

# Outline

- Background
- Evolution of Datasets
- Evolution of Models
- Reflection

# Background: "multi-hop"

- Compositional → interpretable
- Put together information from multiple…

Sentences/documents:

- Reading comprehension
- Open-domain QA

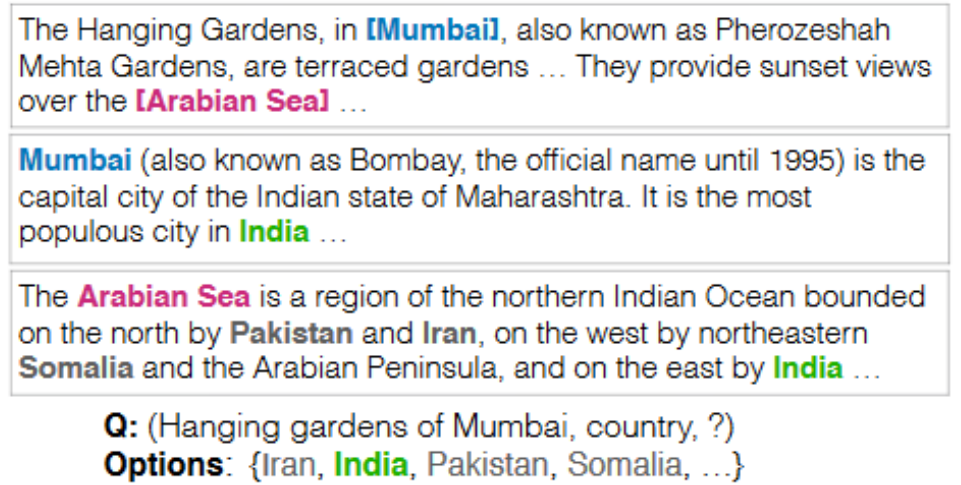



Figure 1: A sample from the WIKIHOP dataset where it is necessary to combine information spread across multiple documents to infer the correct answer.

Figure from Welbl et al., 2018

# Background: "multi-hop"

- Compositional → interpretable
- Put together information from multiple…

Facts in a knowledge graph:

- Knowledge graph completion

Both:

- Open-domain/commonsense QA

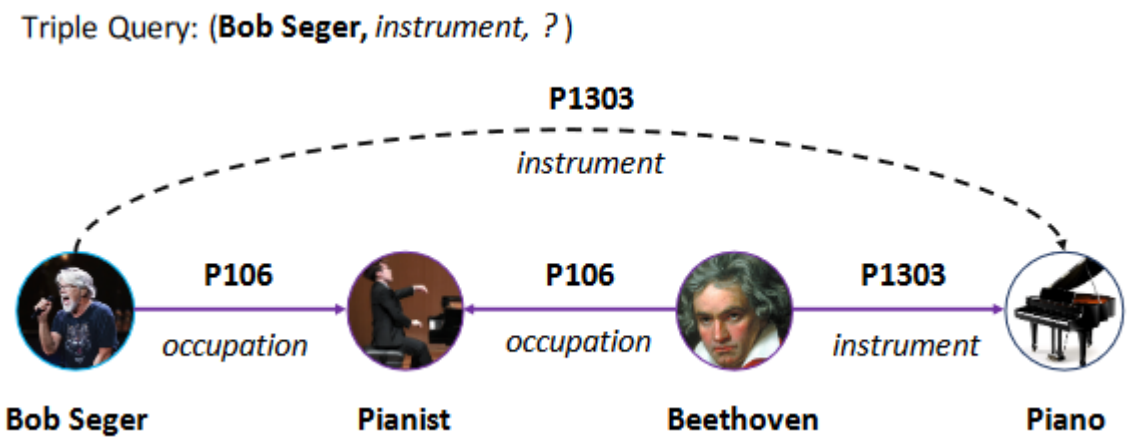



Figure from Lv et al., 2021

# Outline

- Background
- Evolution of Datasets
  - HotPotQA
  - 3 EMNLP papers
- Evolution of Models
- Reflection

Green: at EMNLP 2021
Gray/faded: not discussed in detail

# Datasets: Motivation

- What was the goal in 2018?

"the [existing] datasets do not actually require a deep understanding. We address this by developing a challenge in which answering a question requires reasoning over multiple sentences." – Khashabi et al., 2018

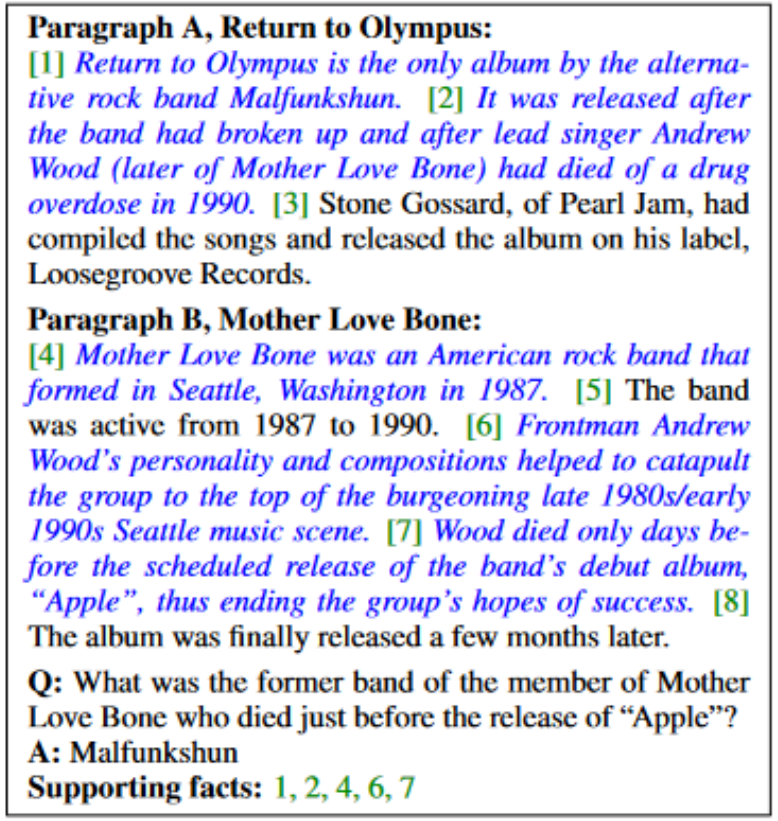
"investigate the limits of existing methods" – Welbl et al., 2018

"Answering complex questions can be addressed by decomposing the question into … simple questions" – Talmor and Berant, 2018

"explainable question answering dataset" – Yang, Qi, Zhang et al., 2018

# Datasets: Multi-hop (in 2018)

- KB-based:
  - QAngaroo (Welbl et al., 2018)
  - ComplexWebQuestions (Talmor and Berant, 2018)
- Documents only:
  - Hotpot-QA (Yang, Qi, Zhang et al., 2018)
  - MultiRC (Khashabi et al., 2018)
  - WorldTree (Jansen et al., 2018), WorldTree V2 (Xie et al., 2020)
- Both:
  - OpenBookQA (Mihaylov et al., 2018)



**Paragraph A, Return to Olympus:**
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

**Paragraph B, Mother Love Bone:**
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

**Q:** What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
**A:** Malfunkshun
**Supporting facts:** 1, 2, 4, 6, 7

Figure 1: An example of the multi-hop questions in HOTPOTQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

Green: at EMNLP 2021
Gray/faded: not discussed in detail

# Datasets: HotpotQA (Yang, Qi, Zhang et al., 2018)

**HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering**

**Zhilin Yang*♠ Peng Qi*♡ Saizheng Zhang*♣**
**Yoshua Bengio♣◇ William W. Cohen† Ruslan Salakhutdinov♠ Christopher D. Manning♡**
♠ Carnegie Mellon University ♡ Stanford University ♣ Mila, Université de Montréal
◇ CIFAR Senior Fellow † Google AI
{zhiliny, rsalakhu}@cs.cmu.edu, {pengqi, manning}@cs.stanford.edu
saizheng.zhang@umontreal.ca, yoshua.bengio@gmail.com, wcohen@google.com

113K Wikipedia-based questions

1. Require reasoning over multiple documents
2. Diverse, not constrained to pre-existing schemas/KBs
3. Strongly supervised: supported facts are provided
4. Factoid comparison task

[1]The name comes from the first three authors' arriving at the main idea during a discussion at a hot pot restaurant.

# HotPotQA: Data Collection

1. Build Wikipedia hyperlink graph
2. Sample paragraph pairs: (A, B)
3. Write questions requiring paragraphs A and B
   - Sometimes ask a comparison question
4. Collect supporting facts

**Paragraph A, Return to Olympus:**
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

**Paragraph B, Mother Love Bone:**
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

**Q:** What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
**A:** Malfunkshun
**Supporting facts:** 1, 2, 4, 6, 7

Figure 1: An example of the multi-hop questions in HotpotQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

# HotPotQA: The Benchmark

1. Categorize questions/turkers into:
   - “easy” (single-hop),
   - “medium” (multi-hop answerable with 2018 QA models)
   - “hard” (the rest)
2. Test:
   - “distractor”: where 2 gold + 8 related paragraphs per question
   - “full wiki”: no paragraphs are provided

# HotPotQA: Dataset Analysis

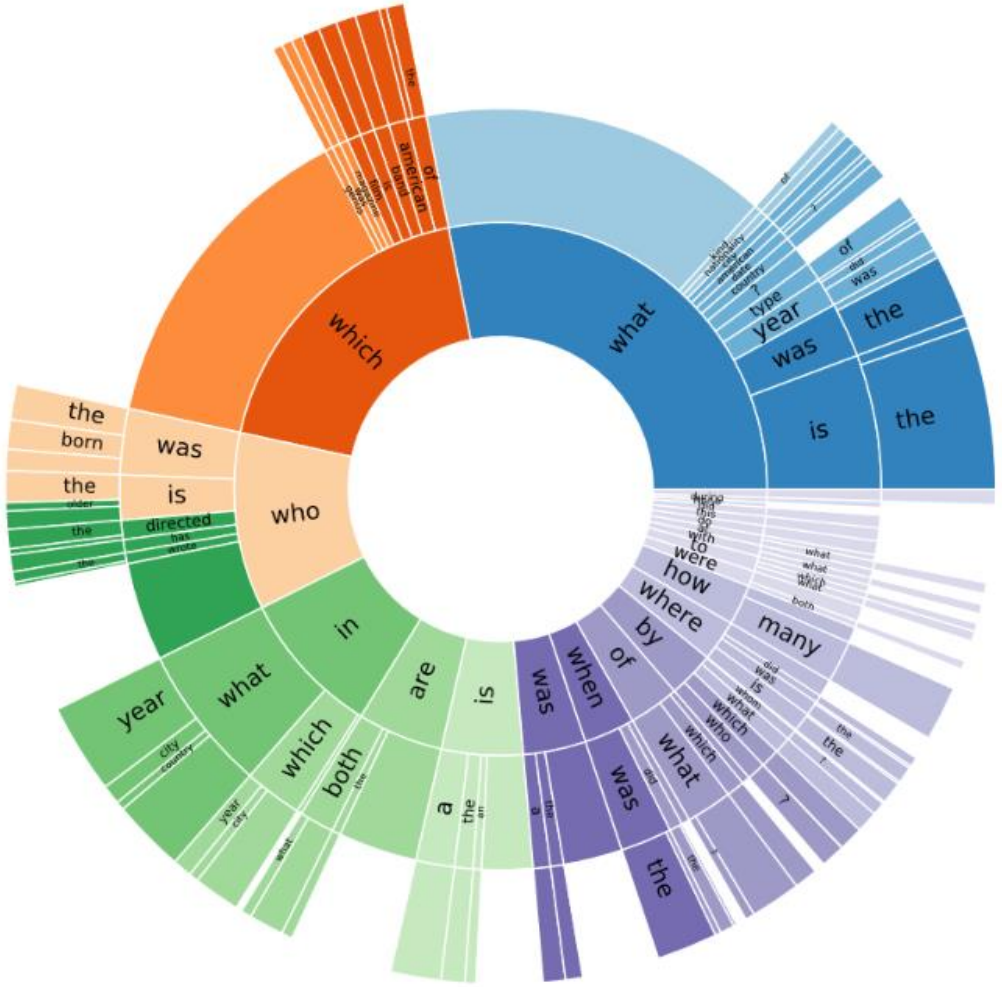



Figure 2: Types of questions covered in HOTPOTQA.

| Answer Type | % | Example(s) |
|---|---|---|
| Person | 30 | King Edward II, Rihanna |
| Group / Org | 13 | Cartoonito, Apalachee |
| Location | 10 | Fort Richardson, California |
| Date | 9 | 10th or even 13th century |
| Number | 8 | 79.92 million, 17 |
| Artwork | 8 | Die schweigsame Frau |
| Yes/No | 6 | - |
| Adjective | 4 | conservative |
| Event | 1 | Prix Benois de la Danse |
| Other proper noun | 6 | Cold War, Laban Movement Analysis |
| Common noun | 5 | comedy, both men and women |

Table 2: Types of answers in HOTPOTQA.

# HotPotQA: Examples

| Reasoning Type | % | Example(s) |
|---|---|---|
| Inferring the *bridge entity* to complete the 2nd-hop question (Type I) | 42 | **Paragraph A:** The 2015 Diamond Head Classic was a college basketball tournament ... *Buddy Hield was named the tournament's MVP*.<br>**Paragraph B:** *Chavano Rainier "Buddy" Hield* is a Bahamian professional basketball player for the **Sacramento Kings** of the NBA...<br>**Q:** Which team does the player named 2015 Diamond Head Classic's MVP play for? |
| Comparing two entities (Comparison) | 27 | **Paragraph A:** LostAlone were a British rock band ... consisted of *Steven Battelle, Alan Williamson, and Mark Gibson*...<br>**Paragraph B:** Guster is an American alternative rock band ... Founding members *Adam Gardner, Ryan Miller, and Brian Rosenworcel* began...<br>**Q:** Did LostAlone and Guster have the same number of members? (**yes**) |
| Locating the **answer entity** by checking multiple properties (Type II) | 15 | **Paragraph A:** Several *current and former members of the Pittsburgh Pirates* ... John Milner, **Dave Parker**, and Rod Scurry...<br>**Paragraph B:** **David Gene Parker**, *nicknamed "The Cobra"*, is an American former player in Major League Baseball...<br>**Q:** Which former member of the Pittsburgh Pirates was nicknamed "The Cobra"? |

# HotPotQA: More details in paper

- Modeling for both distractor and full-wiki setting
- Using supporting facts as strong supervision
- Human performance
- Examples

# Datasets: What's new?

- Addressing limitations of multi-hop evaluation:

Answering Open-Domain Questions of Varying Reasoning Steps from Text (Qi, Lee, Sido et al., 2021)

- Addressing multi-hop as "explanation"

On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings (Jansen et al., 2021)

- Knowledge graph reasoning:

Is Multi-Hop Reasoning Really Explainable? Towards Benchmarking Reasoning Interpretability (Lv et al., 2021)

Green: at EMNLP 2021
Gray/faded: not discussed in detail

# Datasets: BeerQA (Qi, Lee, Sido et al., 2021) @ EMNLP

**Answering Open-Domain Questions of Varying Reasoning Steps from Text**

**Peng Qi*♠♡ Haejun Lee*♣ Oghenetegiri "TG" Sido*♠ Christopher D. Manning♠**
♠ Computer Science Department, Stanford University
♡ JD AI Research
♣ Samsung Research
{pengqi, osido, manning}@cs.stanford.edu, haejun82.lee@samsung.com

- Revisit unrealistic assumptions of:
  - Knowing how many steps/hops are required
  - Using KBs/web links
- New dataset:
  - Include questions that involve 3 links

# BeerQA: How many hops is “multi”?

- In prior datasets, one or two documents (or paragraphs)
- In real-world, it’s not easy to know whether it’s one or two (or more), e.g.

  *In which U.S. state was Facebook founded?*

- Hyperlinks aren’t always available during retrieval

# BeerQA: New benchmark

- Varying hops:
  - One-hop: SQuAD Open
  - Two hop: HotPotQA multi-hop
  - 3+ hop: New collection of questions
- Unified Wikipedia (filter out unanswerable questions)

# BeerQA: 3+ hop dataset

- 50-100 question templates covering diverse topics
- 10-20 examples/question

| # of Documents to answer the question | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| # of questions | 495 | 17 | 8 | 0 | 9 | 1 |

Table 9: Distribution of reasoning steps for questions in Three+ Hop Challenge Set.

| Reasoning Type | % |
|---|---|
| Comparison | 25.6 |
| Bridge-Comparison | 25.3 |
| Bridge | 49.1 |

Table 10: Reasoning types required for Three+ Hop Challenge Set.

| Answer Type | % | Example(s) |
|---|---|---|
| Person | 29 | Kate Elizabeth Winslet |
| Number | 20 | 388,072, 5.5 million |
| Yes / No | 15 | — |
| Group / Org | 11 | CNN |
| Date | 8 | March 28, 1930 |
| Other Proper Noun | 7 | Boeing 747-400 |
| Creative Work | 5 | “California Dreams” |
| Location | 4 | New York City |
| Common Noun | 1 | comedy-drama |

Table 11: Types of answers in Three+ Hop Challenge Set. These statistics are based on 100 randomly sampled examples.

# BeerQA: Example

**Question:** How many counties are on the island that is home to the fictional setting of the novel in which Daisy Buchanan is a supporting character?

**Wikipedia Page 1:** ***Daisy Buchanan***

Daisy Fay Buchanan is a fictional character in F. Scott Fitzgerald's magnum opus "The Great Gatsby" (1925)...

**Wikipedia Page 2:** ***The Great Gatsby***

The Great Gatsby is a 1925 novel ... that follows a cast of characters living in the fictional town of West Egg on prosperous Long Island ...

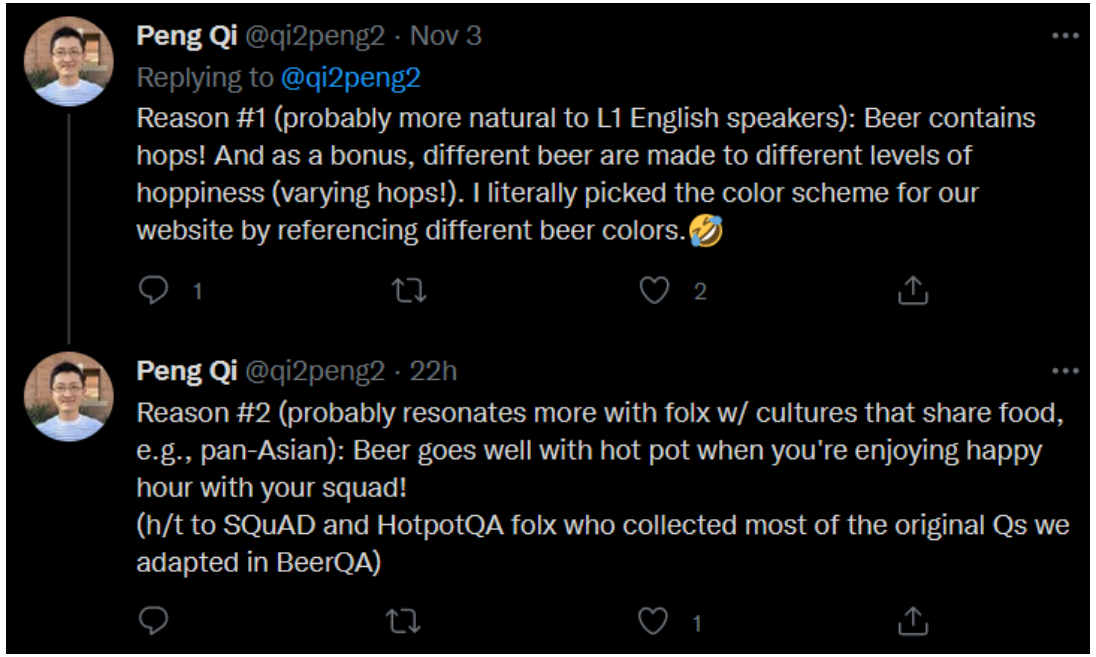
**Wikipedia Page 3:** ***Long Island***

The Long Island ... comprises four counties in the U.S. state of New York: Kings and Queens ... to the west; and Nassau and Suffolk to the east...

**Answer:** four

# BeerQA: More details in paper

- Novel model for Open-domain QA based on iteratively retrieving, reading, and reranking (IRRR)
- Competitive performance on SQuAD and HotPotQA = Strong benchmark for BeerQA
- Why is it called BeerQA?
  - Twitter thread h/t Marc



# Datasets: What's new?

- Addressing limitations of multi-hop evaluation:

Answering Open-Domain Questions of Varying Reasoning Steps from Text (Qi, Lee, Sido et al., 2021)

- Addressing multi-hop as "explanation"

On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings (Jansen et al., 2021)

- Knowledge graph reasoning:

Is Multi-Hop Reasoning Really Explainable? Towards Benchmarking Reasoning Interpretability (Lv et al., 2021)

Green: at EMNLP 2021
Gray/faded: not discussed in detail

# Datasets: Extending WorldTree V2 (Jensen et al., 2021) @ EMNLP

**On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings**

**Peter A. Jansen** and **Kelly Smith** and **Dan Moreno** and **Huitzilin Ortiz**
University of Arizona, USA
pajansen@arizona.edu

"a desirable consequence is that the facts used to assemble this chain-of-reasoning can then be taken as an interpretable record of that reasoning, as well as a human-readable explanation for why it is correct"

# WorldTreeV2+: Motivation

- Prior multi-hop work isn't very "multi"
- Multiple reference problem for explanations
  - Automatic vs. expert evaluation?
- Goal: formalize evaluation by examining **relevance** and **completeness** of explanation

**Question:** When trees are cleared from the land, what will most likely occur?
**Answer:** Soil Erosion

**Gold Explanation**

A tree is a kind of plant.
Roots are a part of a plant.
In the soil erosion process, plant roots are an inhibitor.
Removing an inhibitor causes that process to happen.

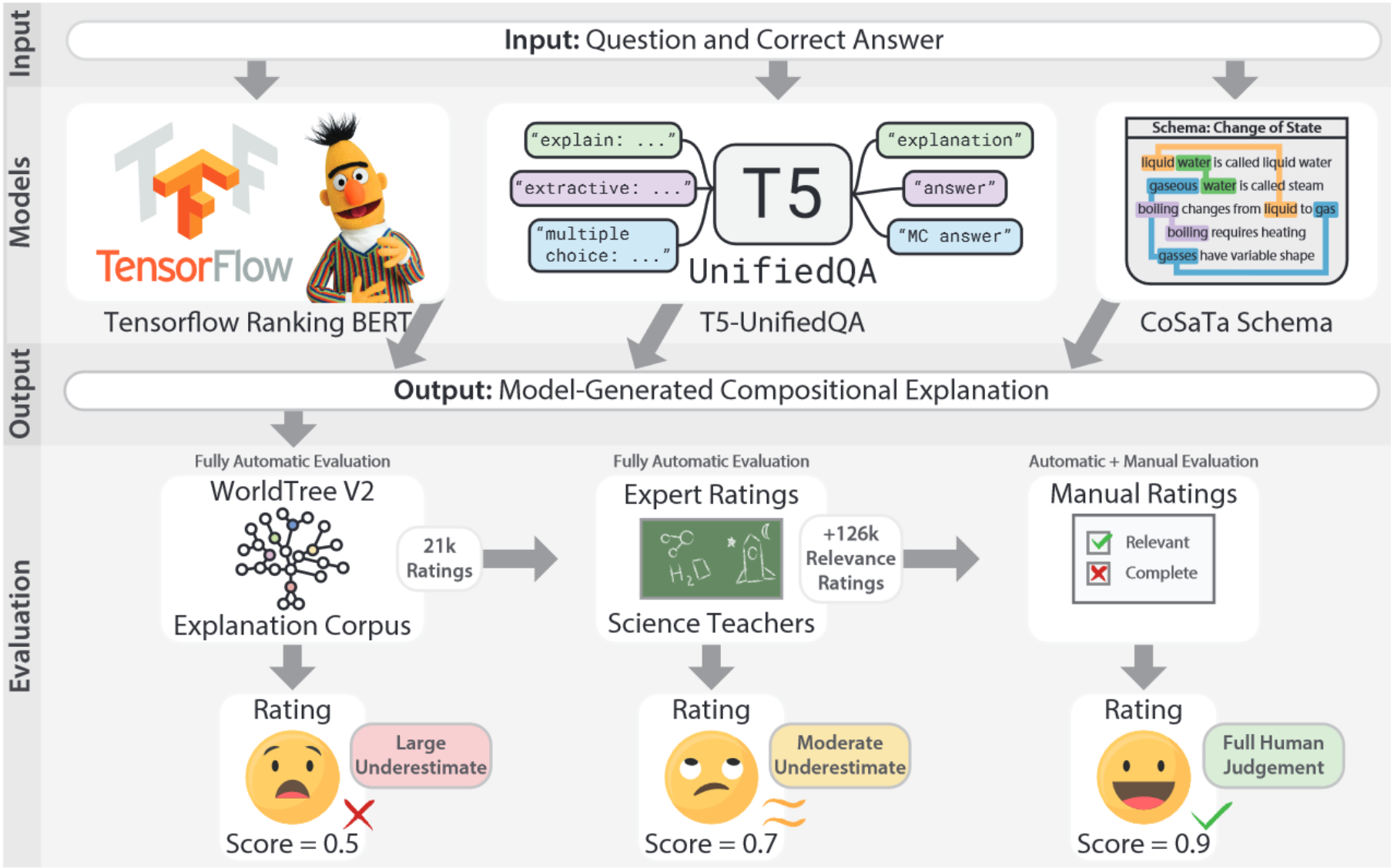
**Model-Generated Explanation**

Soil erosion is when wind/water move soil.
Tree roots decrease soil erosion.
As deforestation increases, soil erosion will increase.
Deforested area is where humans cut down trees.
Clearing a forest means cutting down trees.

Automated Evaluation ✗

Expert Evaluation ✓

# WorldTreeV2+: Evaluation Extension



# WorldTreeV2+: Rating 126K facts for relevance

- WorldTree V2 contains 4.4K questions, 9K facts → 40M ratings if exhaustive
- Use BERT and RoBERTa to create a shortlist of facts per question
  - 28.9 facts/question
- Annotate a shortlist of 126K facts with experts (8-20 years of science teacher experience)
- Interannotator agreement $\kappa = 0.46$
- Final score average + round up

# WorldTreeV2+: Rating 126K facts for relevance

- ~18 new relevant facts/question

| TR | Label | Description |
|---|---|---|
| 3 | Core | Facts that directly address the core topic the question is testing. |
| 2 | Important | Key knowledge supporting the core facts or grounding core knowledge in examples the question uses. |
| 1 | Extra Detail | Facts that (a) when included, add extra detail to the explanation, but (b) when missing, do not exclude important details from the explanation. |
| 0 | Irrelevant | Facts not relevant to the question. |

Table 2: 4-point Relevance Rating Scheme

**Q:** Burning fossil fuels adds pollutants like sulphur into the air. This pollution contributes to:
**A:** acid rain

| TR | Gold | Fact |
|---|---|---|
| 3 | * | Burning fossil fuels releases sulfur dioxide into the atmosphere. |
| 3 | * | Emitting sulfur dioxide causes acid rain. |
| 2 | | Burning fossil fuels causes pollution. |
| 2 | * | Emission is when something is added to. the atmosphere. |
| 2 | | Gasses from burning oil and coal that dissolve in water in the atmosphere cause acid rain. |
| 2 | | As the amount of sulphur gas in the atmosphere increases, the PH of rain will decrease. |
| 1 | | Acid rain negatively impacts water quality. |
| 1 | | Coal is a kind of fossil fuel. |
| 0 | | The air contains carbon dioxide. |
| 0 | | Oil is a kind of pollutant. |

# WorldTreeV2+: Evaluation of whole explanations

- Models: T5-UnifiedQA (generative), TFR-BERT (reranking), CoSaTa (schema)
- **Relevance**: proportion that have non-zero relevance score
- **Completeness**:
  - Automated: recall of gold explanation
  - Expert-based (B): binary metric of 1 if all facts were rated 2 or 3
- F1 between relevance and completeness

# WorldTreeV2+: Automated Evaluation

- **Relevance**: proportion that have non-zero relevance score
- **Completeness**:
  - Recall of gold explanation
  - Expert-based (B): binary metric of 1 if all facts were rated 2 or 3
- See Table 5 for exact numbers
  - TFR-BERT (reranking) does best
  - CoSaTa (schema) has high relevance (precision)

# WorldTreeV2+: Manual Evaluation

- **Relevance**: proportion that have non-zero relevance score
- **Completeness**:
  - Manually rated for 50 dev questions

| | Automatic Analysis | | | | | Manual Analysis | | | Underestimate ($\Delta$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pattern Scoring Method | Rel | Comp | $F1^{ex}$ | $\text{Comp}_B$ | $F1^{ex}_B$ | Rel | $\text{Comp}_B$ | $F1^{ex}_B$ | Rel | $\text{Comp}_B$ | $F1^{ex}_B$ |
| T5-UQA-3B$_{CORE}$ | 0.53 | 0.36 | 0.43 | 0.10 | 0.17 | 0.82 | 0.44 | 0.57 | **+0.29** | +0.34 | **+0.40** |
| TFR-BERT | 0.72 | **0.59** | **0.65** | **0.36** | **0.48** | **0.93** | **0.72** | **0.81** | +0.21 | **+0.36** | +0.33 |
| Schema (3 Schemas) | **0.74** | 0.46 | 0.57 | 0.21 | 0.33 | 0.79 | 0.44 | 0.57 | +0.05 | +0.23 | +0.24 |

# WorldTreeV2+: Challenges

- Relevance: model performance is undercounted
  - Facts aren't always relevant in isolation
- Completeness: Major challenge as explanations get larger
- Automated metrics: undercounting disproportionately affects model comparison

# Datasets: What's new?

- Addressing limitations of multi-hop evaluation:

Answering Open-Domain Questions of Varying Reasoning Steps from Text (Qi, Lee, Sido et al., 2021)

- Addressing multi-hop as "explanation"

On the Challenges of Evaluating Compositional Explanations in Multi-Hop Inference: Relevance, Completeness, and Expert Ratings (Jansen et al., 2021)

- Knowledge graph reasoning:

Is Multi-Hop Reasoning Really Explainable? Towards Benchmarking Reasoning Interpretability (Lv et al., 2021)

Green: at EMNLP 2021
Gray/faded: not discussed in detail

# Datasets: BIMR (Lv et al., 2021) @ EMNLP

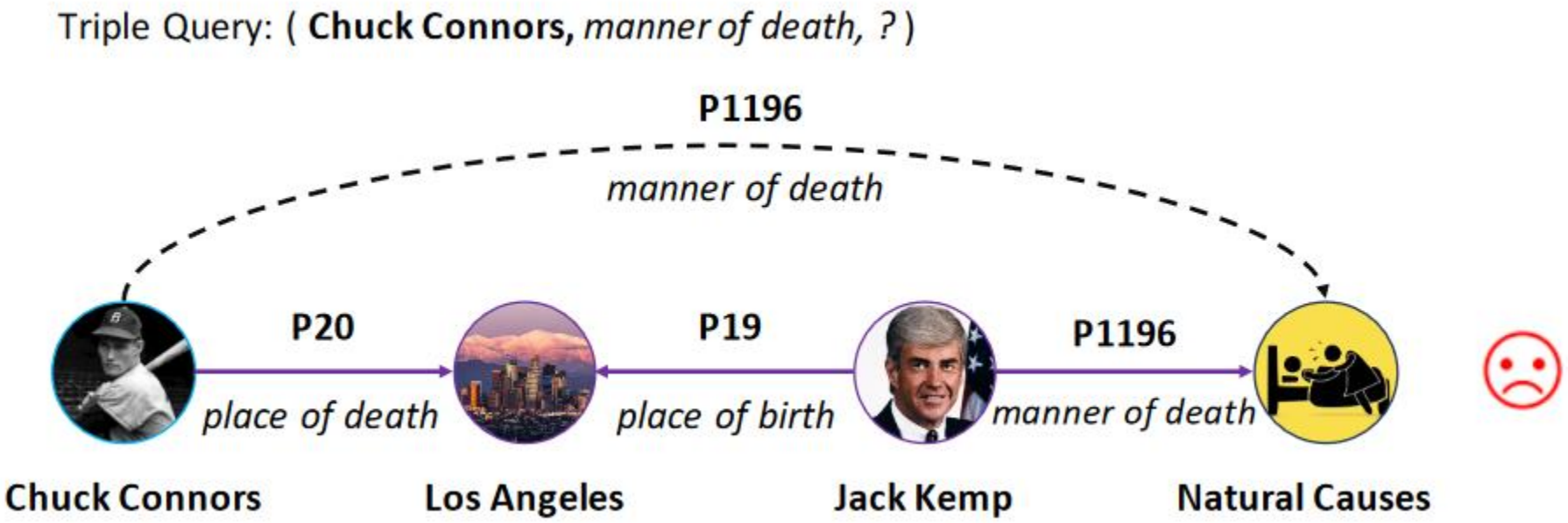
**Is Multi-Hop Reasoning Really Explainable?**
**Towards Benchmarking Reasoning Interpretability**

**Xin Lv**[1,2], **Yixin Cao**[3], **Lei Hou**[1,2*], **Juanzi Li**[1,2]
**Zhiyuan Liu**[1,2], **Yichi Zhang**[4], **Zelin Dai**[4]
[1]Department of Computer Science and Technology, BNRist
[2]KIRC, Institute for Artificial Intelligence, Tsinghua University, Beijing 100084, China
[3]Nanyang Technological University, Singapore
[4]Alibaba Group, Hangzhou, China
lv-x18@mails.tsinghua.edu.cn, yixin.cao@ntu.edu.sg
{houlei,lijuanzi,liuzy}@tsinghua.edu.cn

"Multi-hop reasoning has been widely studied in recent years to obtain more interpretable link prediction"

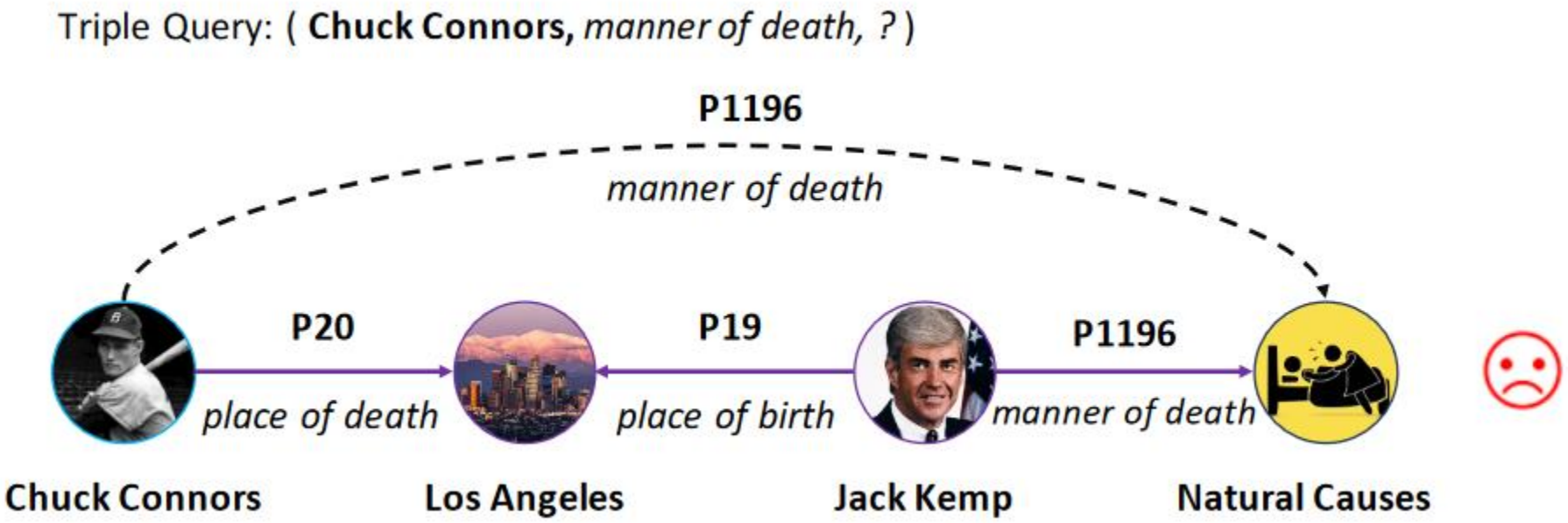
# Datasets: BIMR (Lv et al., 2021) @ EMNLP

"Multi-hop reasoning has been widely studied in recent years to obtain more interpretable link prediction"



# BIMR: Benchmark to detect the Interpretability of Multi-hop Reasoning

- In prior multi-hop reasoning (KG) models, any path found is considered a reasonable explanation
  - In Lin et al., 2018 model, 60% of the paths are unreasonable



# BIMR: Benchmark to detect the Interpretability of Multi-hop Reasoning

- In prior multi-hop reasoning (KG) models, any path found is considered a reasonable explanation
  - In Lin et al., 2018 model, 60% of the paths are unreasonable

- **Goal**: quantitative evaluation of interpretability

# BIMR: KG Background

- KG = {E, R, T} which is a set of entities (E), relations (R), and triples T = {(h, r, t)} where a **h**ead has **r**elation with **t**ail, e.g.

  (Bob Seger, Occupation, Pianist)

- Multi-hop reasoning: given (h, r, ?) and a KG
  - Find t
  - Return a path (h, r, t) = $(h, r_1, e_1) \rightarrow (e_1, r_2, e_2) \rightarrow \ldots \rightarrow (e_{n-1}, r_n, t)$

# BIMR: Benchmark

- Curate WD15K based on Wikidata and FB15K-237 (Freebase)
  - Some filtering to keep entities in both KG and with common relations
  - 15.8K total entities, 182 relations, 177K triples

# BIMR: Evaluation Framework

- Path recall (PR): ~recall
  - For how many triples does a model find any path at all?
- Local Interpretability (LI): ~precision
  - Are the found paths reasonable (average “interpretability score” of found paths)?
- Global Interpretability (GI): PR * LI
  - Average “interpretability score” for each triple in test set (0 if no path)

# BIMR: Interpretability Score

1. Find all pair shortest paths (16M paths)

$$(h, r, t) \leftarrow (h, r_1, e_1) \wedge (e_1, r_2, e_2) \wedge (e_2, r_3, t),$$

2. Convert paths to relation sequences, or *rules* (96K)

$$r(X, Y) \leftarrow r_1(X, A_1) \wedge r_2(A_1, A_2) \wedge r_3(A_2, Y).$$

3. Annotate 96K rules for interpretability
   - First, 36K mined and scored using SOTA rule mining method
   - Reannotate high-confidence rules (15K), treat all low-confidence and long-tail rules the same

| | H Rules | L Rules | O Rules |
|---|---|---|---|
| Criteria | $f \in \mathcal{F}^A \wedge c(f) \geq 0.01$ | $f \in \mathcal{F}^A \wedge c(f) < 0.01$ | $f \notin \mathcal{F}^A$ |
| # Rules | 15,458 | 5,534 | 75,027 |
| # Paths | 14.8M | 0.7M | 0.8M |
| Avg Score | 0.216 | 0.005 | 0.069 |

# BIMR: Manual annotation

| | | |
|---|---|---|
| | **Rule:** $cast\ member(X, Y) \leftarrow producer(X, A_1) \wedge spouse(A_1, Y)$ | score: 0.0 |
| 1 | (*Veer-Zaara*, *cast member*, *Rani Mukherjee*) ← (*Veer-Zaara*, *producer*, *Aditya Chopra*) ∧ (*Aditya Chopra*, *spouse*, *Rani Mukherjee*) | score: 0.0 |
| 2 | (*Victor Victoria*, *cast member*, *Julie Andrews*) ← (*Victor Victoria*, *producer*, *Blake Edwards*) ∧ (*Blake Edwards*, *spouse*, *Julie Andrews*) | score: 0.0 |
| 3 | (*The Two Tower*, *cast member*, *Peter Jackson*) ← (*The Two Tower*, *producer*, *Fran Walsh*) ∧ (*Fran Walsh*, *spouse*, *Peter Jackson*) | score: 0.0 |
| 4 | (*10*, *cast member*, *Julie Andrews*) ← (*10*, *producer*, *Blake Edwards*) ∧ (*Blake Edwards*, *pouse*, *Julie Andrews*) | score: 0.0 |
| | **Rule:** $instrument(X, Y) \leftarrow occupation(X, A_1) \wedge uses(A_1, Y)$ | score: 0.9 |
| 1 | (*Sheryl Crow*, *instrument*, *guitar*) ← (*Sheryl Crow*, *occupation*, *guitarist*) ∧ (*guitarist*, *uses*, *guitar*) | score: 1.0 |
| 2 | (*Tom Waits*, *instrument*, *piano*) ← (*Tom Waits*, *occupation*, *pianist*) ∧ (*pianist*, *uses*, *piano*) | score: 1.0 |
| ... | ... | ... |
| 9 | (*Carly Simon*, *instrument*, *guitar*) ← (*Carly Simon*, *occupation*, *guitarist*) ∧ (*guitarist*, *uses*, *guitar*) | score: 0.5 |
| 10 | (*Bruce Hornsby*, *instrument*, *piano*) ← (*Bruce Hornsby*, *occupation*, *pianist*) ∧ (*pianist*, *uses*, *piano*) | score: 1.0 |

# BIMR: Findings

- Compare: 5 Rule-based models and 9 multi-hop reasoning models (see details in paper)
- Current multi-hop reasoning models have some interpretability
  - They lag behind rule-based models
- Upper bound is much higher for all models
- Manually annotated scores > automatic interpretability scores

# Datasets: What's new? Takeaways

- Multi-hop isn't hoppy enough
  → BeerQA
- Compositional "reasoning" isn't tested
  → WorldTreeV2+, BeerQA
- Spurious explanations hurt interpretability
  → BIMR, WorldTreeV2+

Related to: "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference" by McCoy et al., 2018

# Outline

- Background
- Evolution of Datasets
- Evolution of Models
  - Select-Answer-Explain
  - 1 EMNLP model
- Reflection

Green: at EMNLP 2021
Gray/faded: not discussed in detail

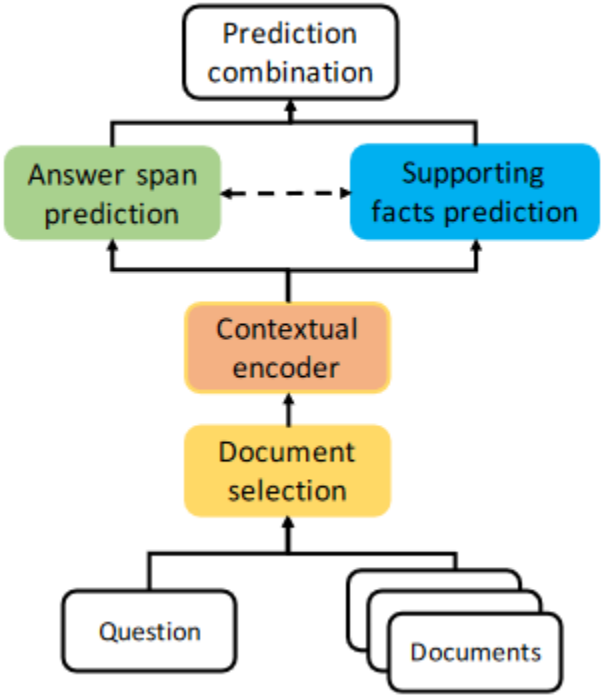
# Models: Which dataset do they work on?

- HotPotQA:
  - Select, Answer and Explain (SAE) (Tu et al., 2020)
  - Summarize-then-Answer: Generating Concise Explanations for Multi-hop Reading Comprehension (Inoue et al., 2021)
  - Answering Open-Domain Questions of Varying Reasoning Steps from Text (Qi, Lee, Sido et al., 2021)
- Commonsense (CORGI)
  - ConversationL mUlti-hop rEasoner (CLUE) (Arabshahi et al., 2021)
- Science QA (OpenBookQA)
  - Chain Guided Reader (CGR) (Xu et al., 2021)

Green: at EMNLP 2021
Gray/faded: not discussed in detail

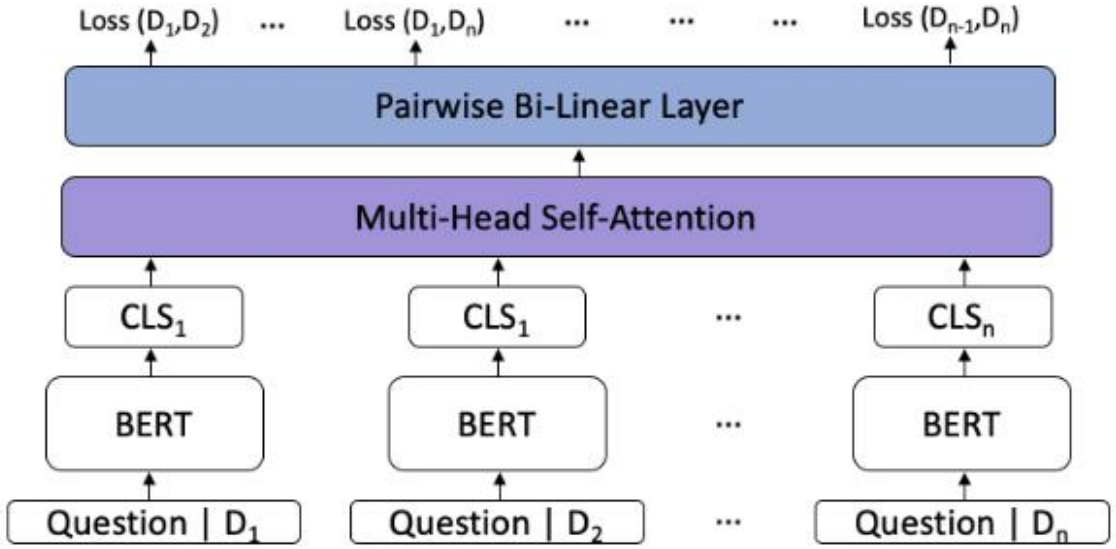
# Models: Select, Answer, Explain (Tu et al., 2020)

- Example of a top-performing public model for HotpotQA for "distractor" setting



# SAE: Document selection

1. Self-attention layer over document embeddings (“CLS”)
2. Pairwise learning-to-rank loss
3. Prioritize:
   - gold document
   - documents containing the answer



# SAE: Answer and explain

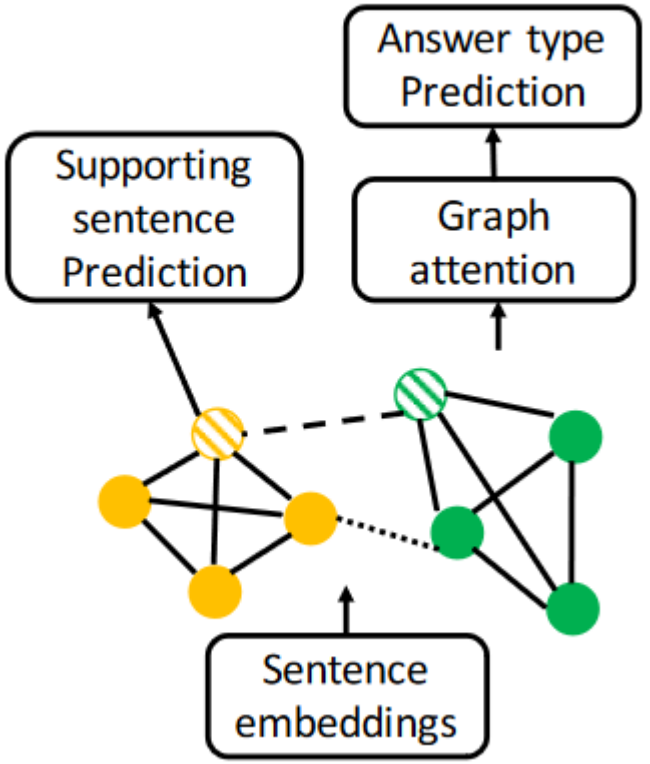
1. Concat + BERT query and paragraphs
2. Multi-task objective for

Answer prediction

- standard QA layer (start/end prediction)

Supporting sentence prediction

- Score each sentence conditioned on predicted answer
- Build graph conv. net over sentence embs
- Message passing strategy to update nodes
- Binary classifier per sentence
- Predict answer type (span or yes/no)



# SAE: Analysis

"Were Scott Derrickson and Ed Wood of the same nationality?"

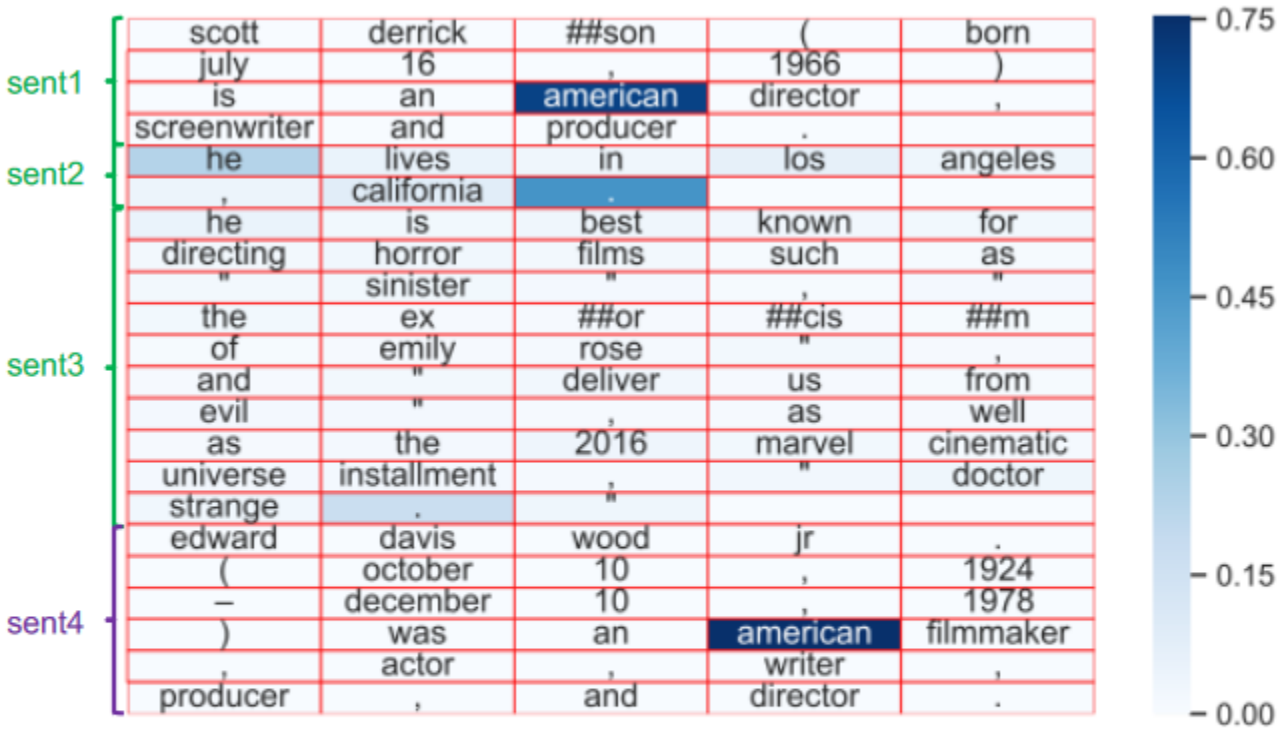



Figure 5: Attention heatmap of a sample from dev set. Each cell is a word piece token returned by BERT. Sentences with different colors are from different documents.

# Models: Which dataset do they work on?

- HotPotQA:
  - Select, Answer and Explain (SAE) (Tu et al., 2020)
  - Summarize-then-Answer: Generating Concise Explanations for Multi-hop Reading Comprehension (Inoue et al., 2021)
  - Answering Open-Domain Questions of Varying Reasoning Steps from Text (Qi, Lee, Sido et al., 2021)
- Commonsense (CORGI)
  - ConversationL mUlti-hop rEasoner (CLUE) (Arabshahi et al., 2021)
- Science QA (OpenBookQA)
  - Chain Guided Reader (CGR) (Xu et al., 2021)

Green: at EMNLP 2021
Gray/faded: not discussed in detail

# Models: Summarize-then-Answer (Inoue et al., 2021) @ EMNLP

**Summarize-then-Answer: Generating Concise Explanations for Multi-hop Reading Comprehension**

**Naoya Inoue**[♣,♠], **Harsh Trivedi**[♣], **Steven Sinha**[♣],
**Niranjan Balasubramanian**[♣], **Kentaro Inui**[♢,♠]
♣ Stony Brook University, ♠ RIKEN
♢ Tohoku University
{ninoue,hjtrivedi,stsinha,niranjan}@cs.stonybrook.edu
inui@tohoku.ac.jp

- Let's abstractively summarize, then use an off-the-shelf QA model

# SuQA: Summarize-then-answer

- Support (evidence) sentences contain *irrelevant* content

**Question**

Charlie Rowe plays Billy Costa in a film based on what novel?

**Paragraphs**

**[P1]** **[1]** The Golden Compass is ~~a 2007 British-American fantasy adventure~~ film based on "Northern Lights", ~~the first novel in Philip Pullman's trilogy "His Dark Materials~~". **[2]** Written and directed by Chris Weitz, it stars Nicole Kidman, Dakota Blue Richards, Daniel Craig, Sam Elliott, Eva Green, and ...
**[P2]** **[1]** Charles John Rowe is an English actor. **[2]** His film roles include ~~Young Tommy in "Never Let Me Go", (...)~~ Billy Costa in "The Golden Compass" , ~~Peter in the SyFy/Sky Movies Peter Pan prequel "Neverland", and recently played Lee Roth on the Fox medical comedy-drama series "Red Band ...~~ ▸ **183 words**

# SuQA: *Concise* definition

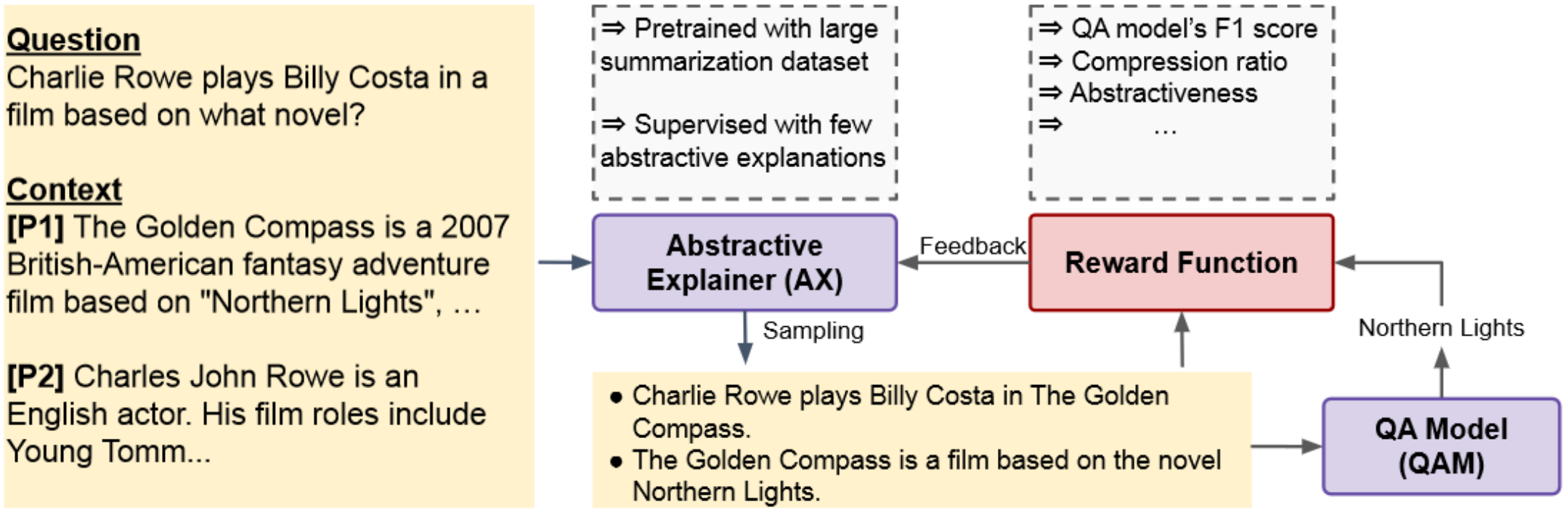
- Minimal  (== relevance?)
- Comprehensible
- Sufficient (== completeness?)

# SuQA: Model

- Abstractive explainer (AX) conditioned on query
  - Concat query + paragraph → BART (pretrained for summarization)
- QA module (QAM)
  - Generation-based QA module

# SuQA: Training

- Pretrain AX on abstractive summarization
- Semi-supervised training with answer as signal



# SuQA: Experiments

- HotpotQA
- Train a document ranker, use top-3 at evaluation
- AX = DistilBART fine-tuned for summarization (CNN/Daily Mail)
- QAM = T5-UnifiedQA-base; frozen
  - In training: UnifiedQA without HotpotQA
  - At test: UnifiedQA finetuned on HotpotQA

# SuQA: Evaluation Metrics

- Conciseness: compression ratio
- Abstractiveness: overlap metrics from reward function
- Sufficiency: 5 annotators on a 3-point Likert scale
  - Due to cost: evaluate 100 gold and 200 generated explanations

- F1: HotpotQA
- XF1: HotpotQA and sufficient

# SuQA: Results

- Abstractive explanations are more concise
  - Compare gold extractive vs. gold abstractive
- Generated abstractive explanations are more concise

| Model | Abs | Cm | Suf† | F1 | XF1† |
|---|---|---|---|---|---|
| QAM w/o AX | 0.0 | 1.0 | - | 64.2 | - |
| Extr (baseline) | 0.3 | 4.2 | 70.0 | **69.4** | 60.5 |
| SuQA-NoRL | 40.1 | 11.2 | 71.5 | 65.6 | 62.6 |
| SuQA | **42.6** | **12.2** | **72.5** | 67.6 | **63.7** |

Table 2: Main results on HotpotQA dev set. †: evaluated on 200 instances with human-judged sufficiency.

# SuQA: Examples

| Insufficiency type | Question | Generated explanation | Gold answer | Freq. |
|---|---|---|---|---|
| No answer span | In which city was this band formed, whose rhythm guitarist featured in "Cupid's Chokehold?" | Cupid's Chokehold is performed by Gym Class Heroes. Fall Out Boy is formed in Wilmette, Illinois. | Chicago | 13 |
| Partially missing | Creed features the boxer who held what WBC title from 2016 to 2017? | Creed (film) features (*missing: the boxer*) Tony Bellew. Tony Bellew held the WBC cruiserweight title from 2016 to 2017. | cruiserweight | 8 |
| Bridge fact missing | Where does the descendant of the Red Setter originate? | James Andrew Hanna is known as Red Setter. Scotch Collie originated from the highland regions of Scotland. *Missing: Scotch Collie is the descendant of Red Setter.* | Scotland | 3 |
| Fact invented | Which game was released first, Icehouse pieces or Kill Doctor Lucky? | Icehouse pieces was released in 1996 (*correct: 1987*). Kill Doctor Lucky was released in 1996. | Icehouse pieces | 1 |
| Dataset flaw | Which Walt Disney film was released earlier, The Rescuers or The Muppets? | The Rescuers was released on June 22, 1977. The Muppets was released in 2011. | The Muppets | 3 |
| Worker error | Does Lucozade pre-date Hires Root Beer? | Hires Root Beer is introduced in 1876. Lucozade is created in 1927. | no | 2 |

Table 6: Manual analysis of 30 insufficient explanations from SuQA.

# SuQA: Limitations

- AX and QAM are trained separately
- Abstractive explainer doesn’t explain the inference process itself

# Outline

- Background
- Evolution of Datasets
- Evolution of Models
- Reflection

# Reflection: multi-hop inference @ EMNLP 2021

- How is “multi-hop reasoning/inference” formally defined?
- What is the purpose of explanation? What needs to be “interpreted”?
  - Model calibration and correctness
  - Simplify fact checking
  - Readability
- How do these interpretability metrics apply to models like T5/GPT-3?