Improving Language Models with Common-Sense Knowledge for Reasoning

(Bill) Yuchen Lin

Advisor: Prof. Xiang Ren

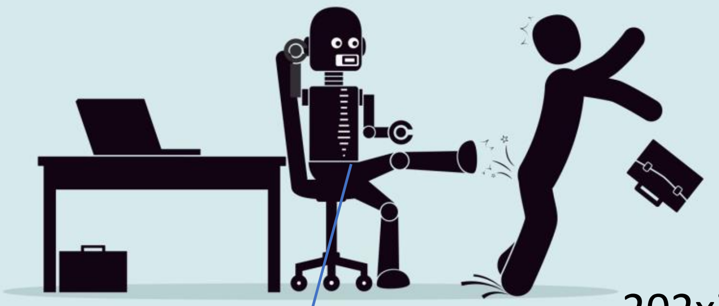
Department of Computer Science

University of Southern California

2021 March

USC University of Southern California

Super-Human Performance in AI?



202x?

Alibaba and Microsoft AI beat human scores on Stanford reading test

Neural networks edged past human scores on the measure of machine reading.

Rob LeFebvre, @roblef
01.15.18 in Personal Computing

10 Comments

937 Shares

Common Sense?

2018 — human-level performance on reading comprehension on SQuAD (Stanford QA dataset)

2017 — super-human performance on speech recognition

2016 — Google neural machine translation

2015 — super-human performance on image captioning

2015 — super-human performance on object recognition

Timeline credits: ACL2020 Tutorial on Commonsense

Teach Machines to Think with Commonsense Knowledge

a) The **common knowledge** that are **shared** among most people in the world.

b) The **reasoning ability** to make decisions in **everyday situations**.

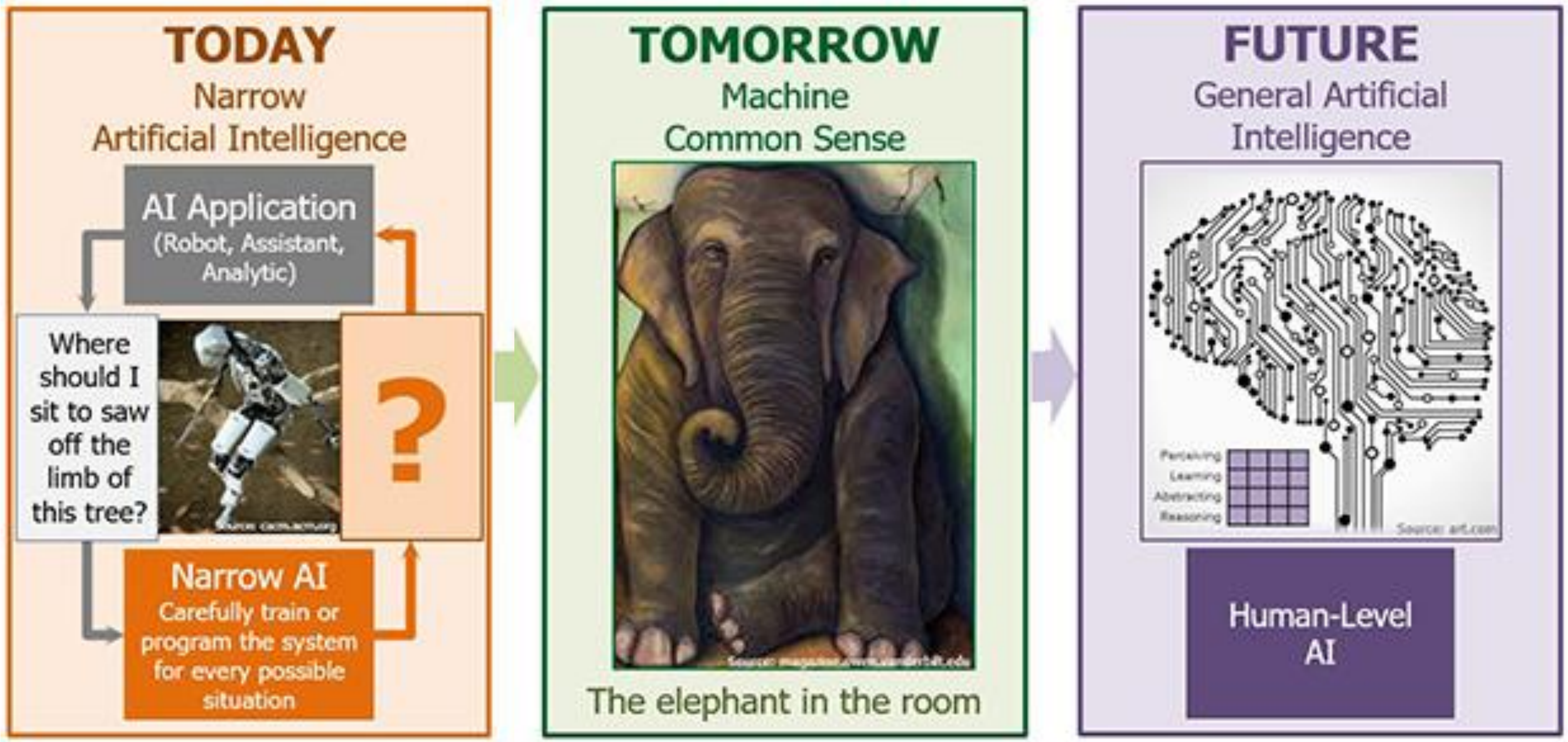


Image Source: DARPA News

What do you fill with ink to write notes on a piece of copy paper?

(A) fountain pen (B) pencil case (C) printer (D) notepad

UNIFIED-QA

State-of-the-art QA Model

Prediction [small, 60 million parameters]: pencil case

Prediction [large, 770 million parameters]: printer

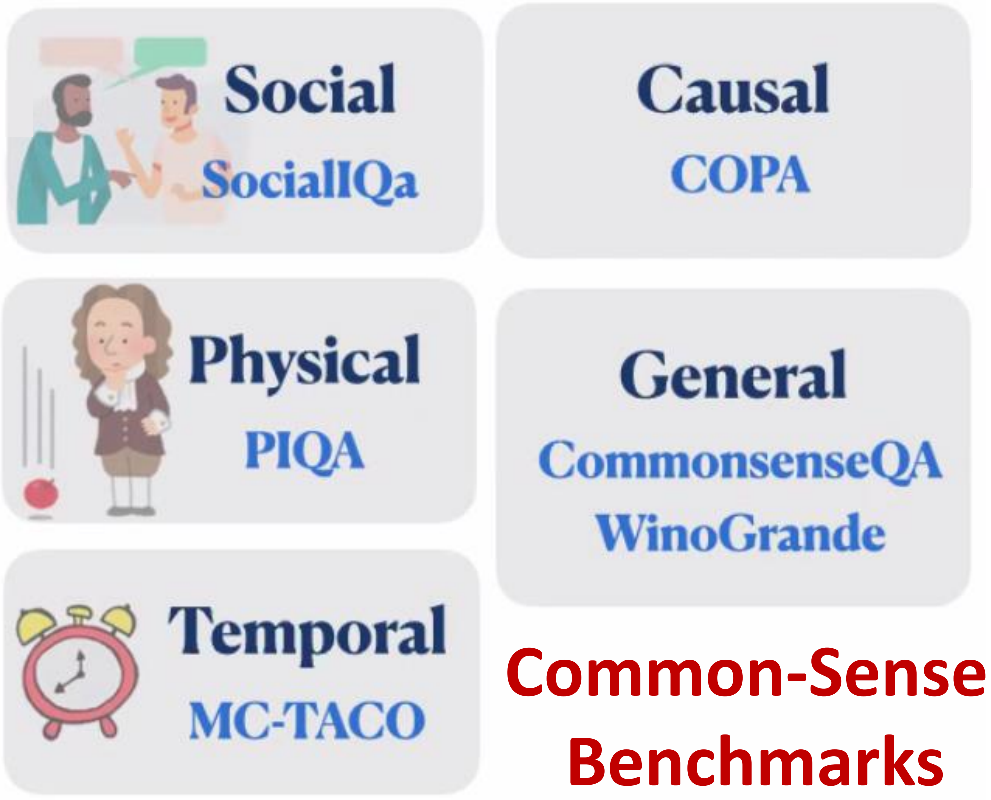
CommonsenseQA (Talmor et al. 2019)

In the school play, Robin played a hero in the struggle to the death with the angry villain.

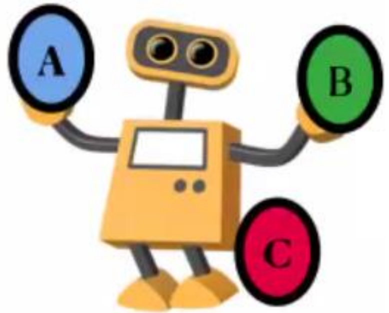
Q How would others feel afterwards?

A (a) sorry for the villain
(b) hopeful that Robin will succeed ✓
(c) like Robin should lose

Social IQA (Sap et al. 2019)



Common-Sense Benchmarks

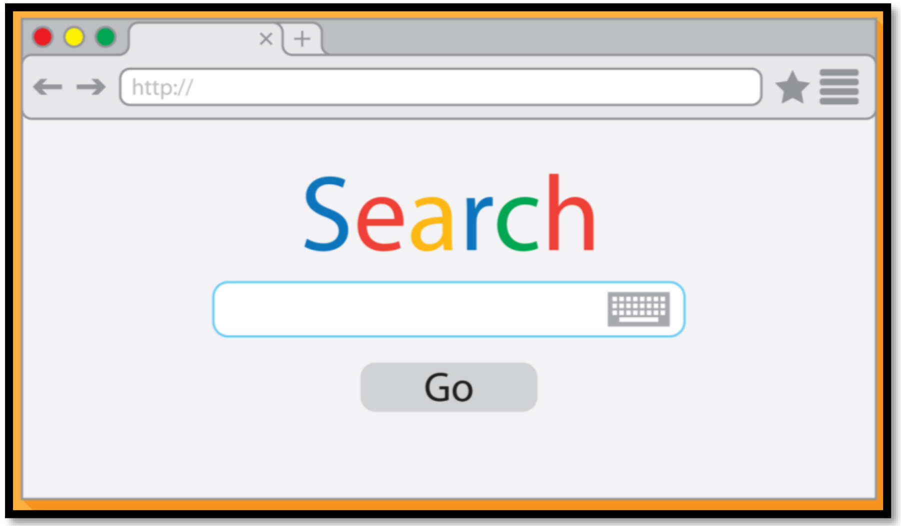


Multiple-Choice Question Answering

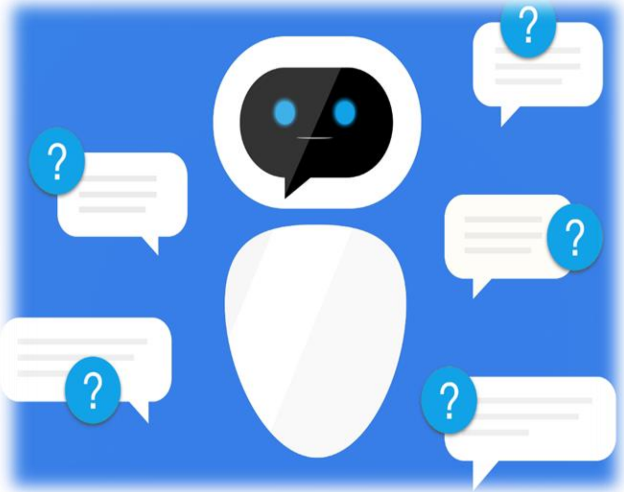
Image credits: Vered Shwartz

Are multiple-choice QA useful in realistic applications?

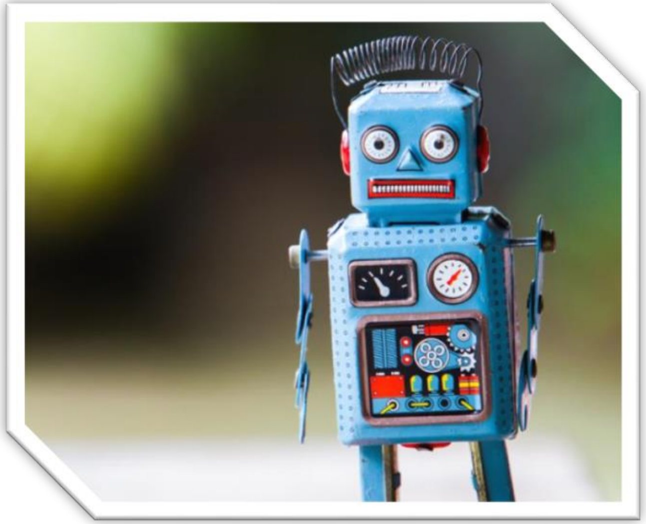
Search Engines



Chatbots



Robots



Common Sense as a Service for practical AI applications.

Real users usually do **NOT** have any ***answer candidates***
when querying commonsense knowledge.

Our Proposal: Open-Ended Commonsense Reasoning

Q: What can help alleviate global warming?

Open-Ended CSR

Input: a question only

A large text corpus of commonsense **facts**

Carbon dioxide is the major greenhouse gas contributing to global warming .

Trees remove *carbon dioxide* from the atmosphere through photosynthesis .

renewable energy, ***tree***, solar battery, …

Output: a ranked list of concepts as answers.

Multiple-Choice/Closed CSR

Input: a question + a few choices

A) air conditioner B) fossil fuel

C) **renewable energy** D) carbon dioxide

?

*Can machines learn to **reason** for such **commonsense** questions?*

Prior Works (1): DrQA --- Retriever + Reader (2017)

Input: a question w/o any candidate choices.

Q: How many of Warsaw's inhabitants spoke Polish in 1933?

WIKIPEDIA
The Free Encyclopedia

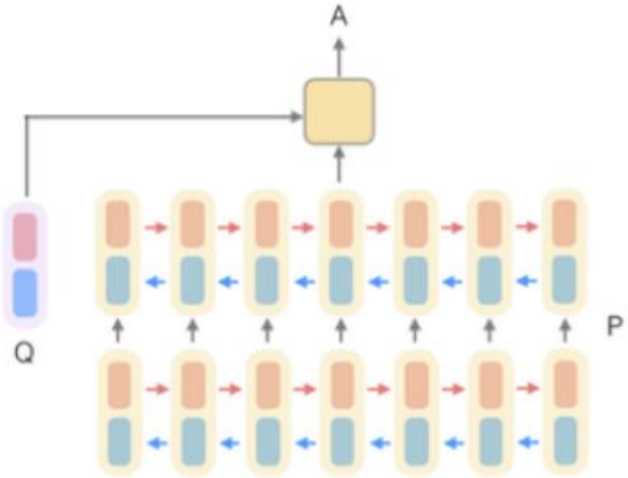
Document
Retriever

I: Question + Passage
O: A span (in the passage)

Document
Reader

Output: a span

833,500



Retriever = Sparse Textual Features + Probabilistic Ranking Alg. (e.g.,BM25)

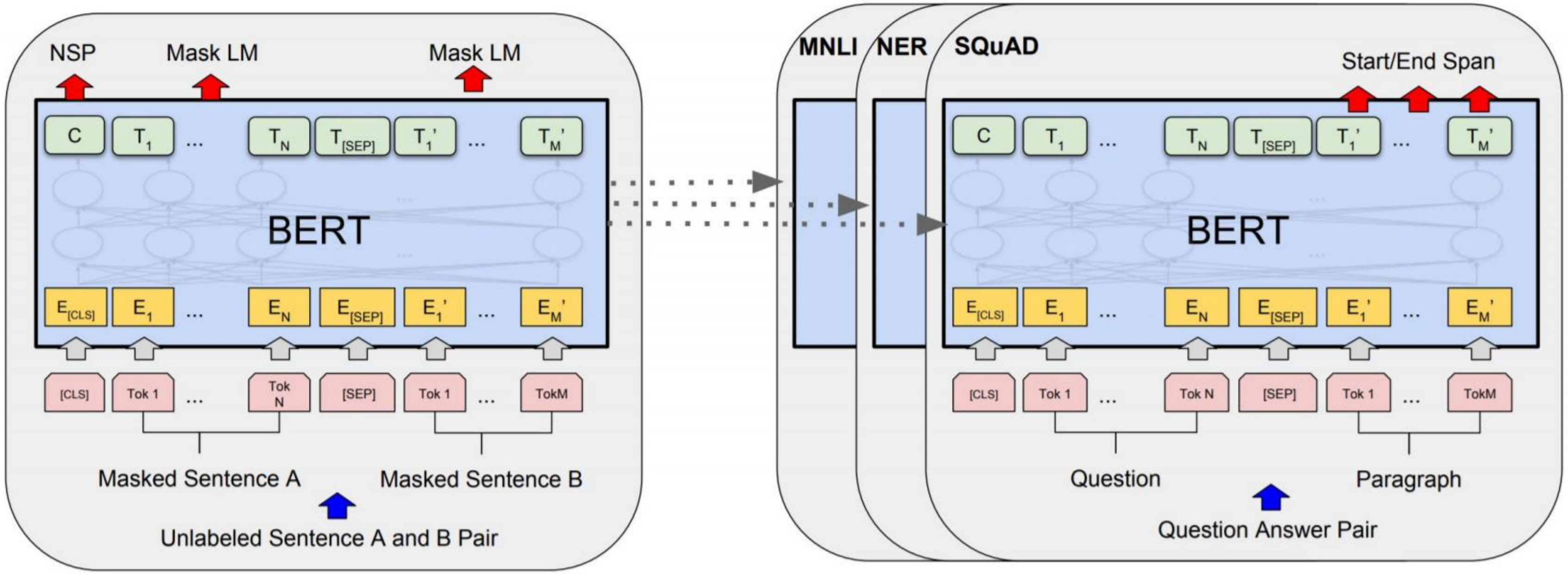
Reader = Machine Learning Methods for **Extracting Answers (e.g., MatchLSTM)**

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions

BERT --- A neural language model. (2018)

```
           store           gallon
             ↑                ↑
the man went to the [MASK] to buy a [MASK] of milk
```

Masked Language Modeling:
Learn a neural model to fill in the blanks
--- *masked words* in a sentence.



Pre-training

WIKIPEDIA
The Free Encyclopedia

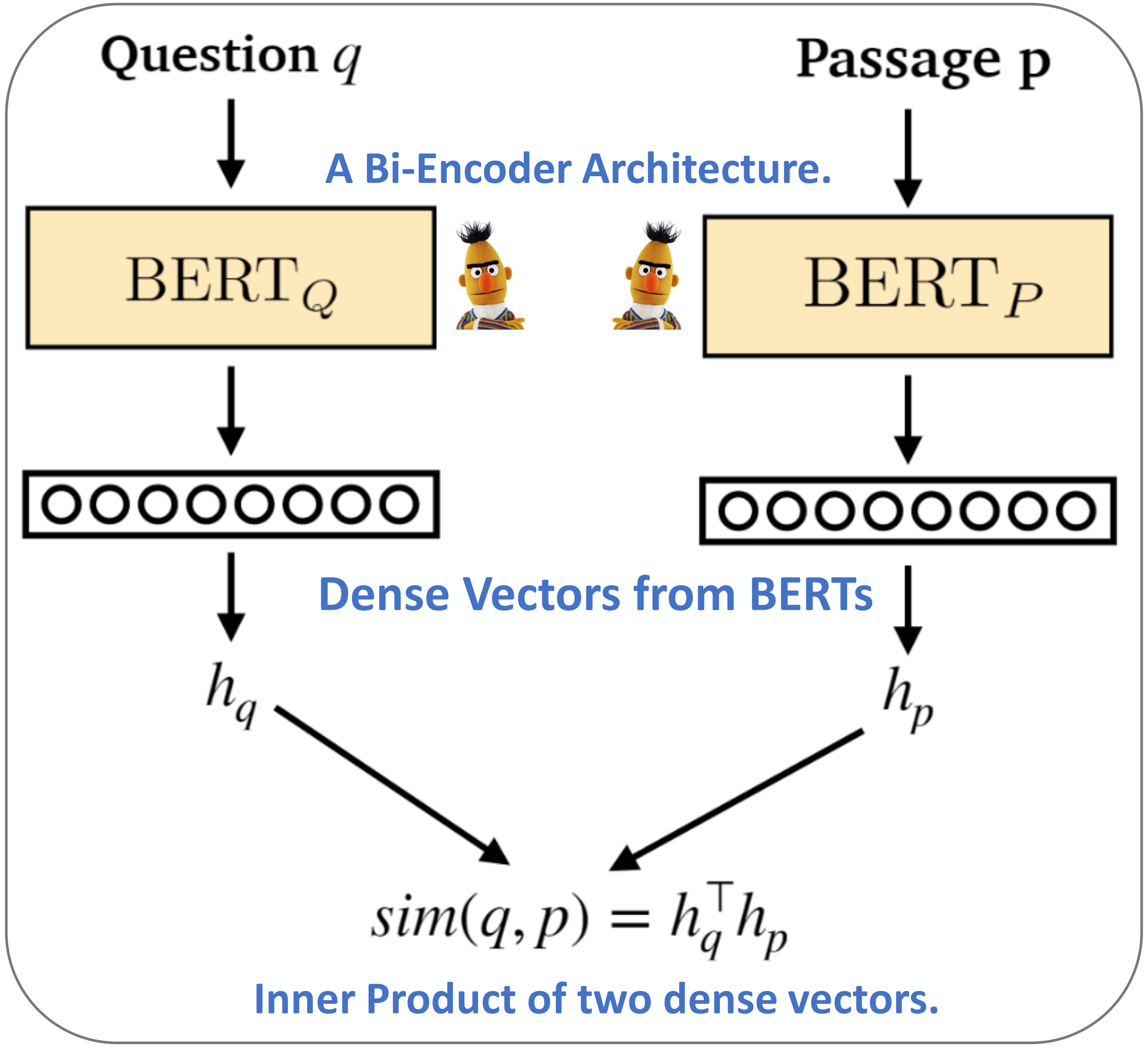
Fine-Tuning

Text Representations for Practical NLP Tasks (e.g., QA)

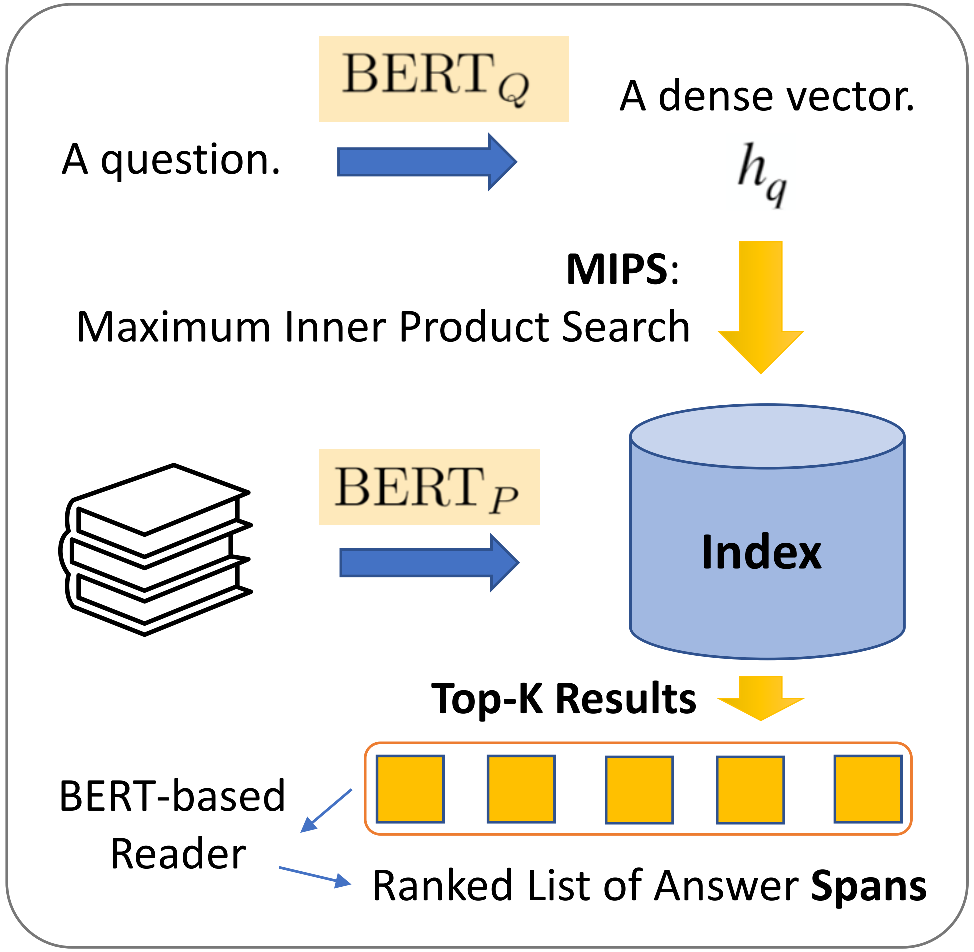
Self-Supervised Pre-Training + Task-Specific Fine-Tuning

Prior Works (2): DPR --- Dense Passage Retriever (2020)

A Trainable Method for Passage Retrieval



MIPS-based Inference Pipeline

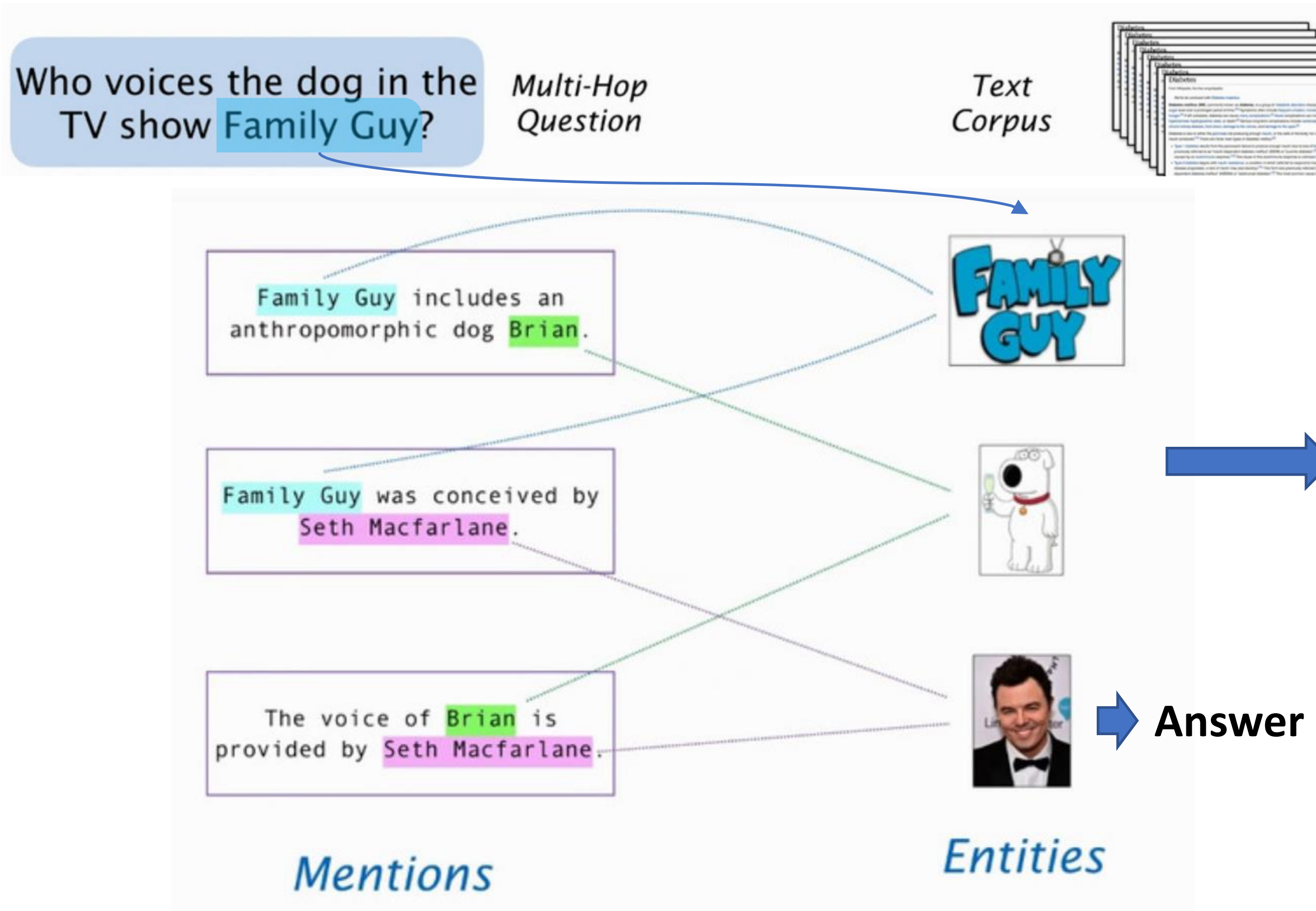


Advantage:
Trainable Rtrv. w/ BERT

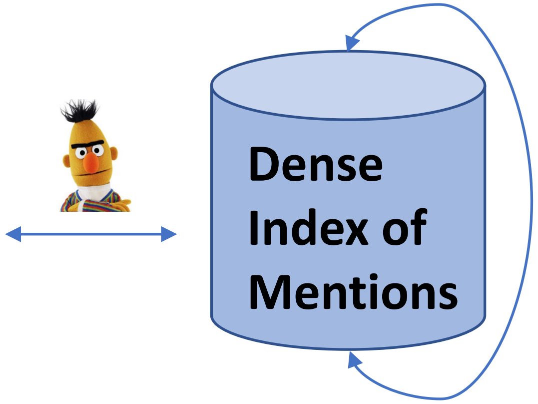
Limitation:
Single-hop Reasoning.

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Prior Works (3): DrKIT --- Differentiable Multi-Hop Reasoning (2020)



A text corpus as a mention-entity bipartite graph.



MIPS + Entity Filtering

Modeling the Entity Following w/ **differentiable operations**.

Disadvantages:

1) Only **mentions**' embedding is modeled.
   Limited contextual info can be used.
2) A **fixed number of steps** within a model.
3) **Inefficient** if the **bipartite graph** is dense.

Dhingra et al. 2020. Differentiable Reasoning over a Virtual Knowledge Base

Why is OpenCSR challenging?

1) Latent Multi-Hop Structures (vs. factoid questions).

Who voices the dog in the TV show Family Guy ?

A multi-hop, factoid question from **HotpotQA**.

q_1= <u>the dog</u> in the TV show Family Guy

q_2= who <u>voices</u> [**q_1. answer**]

Clear, **explicit** hints for querying **evident relations** between **named entities**.

What can help alleviate global warming?

q_1= what <u>*contributes to*</u> global warming

q_2= what <u>*removes*</u> [**q_1. answer**]

Latent, **implicit** hints for querying **complex relations** between **concepts**.

2) Very Large Search Space (vs. multiple-choice setting).

3) Much Dense Entity Links (vs. Wikipedia entities).

Formulating the task of OpenCSR

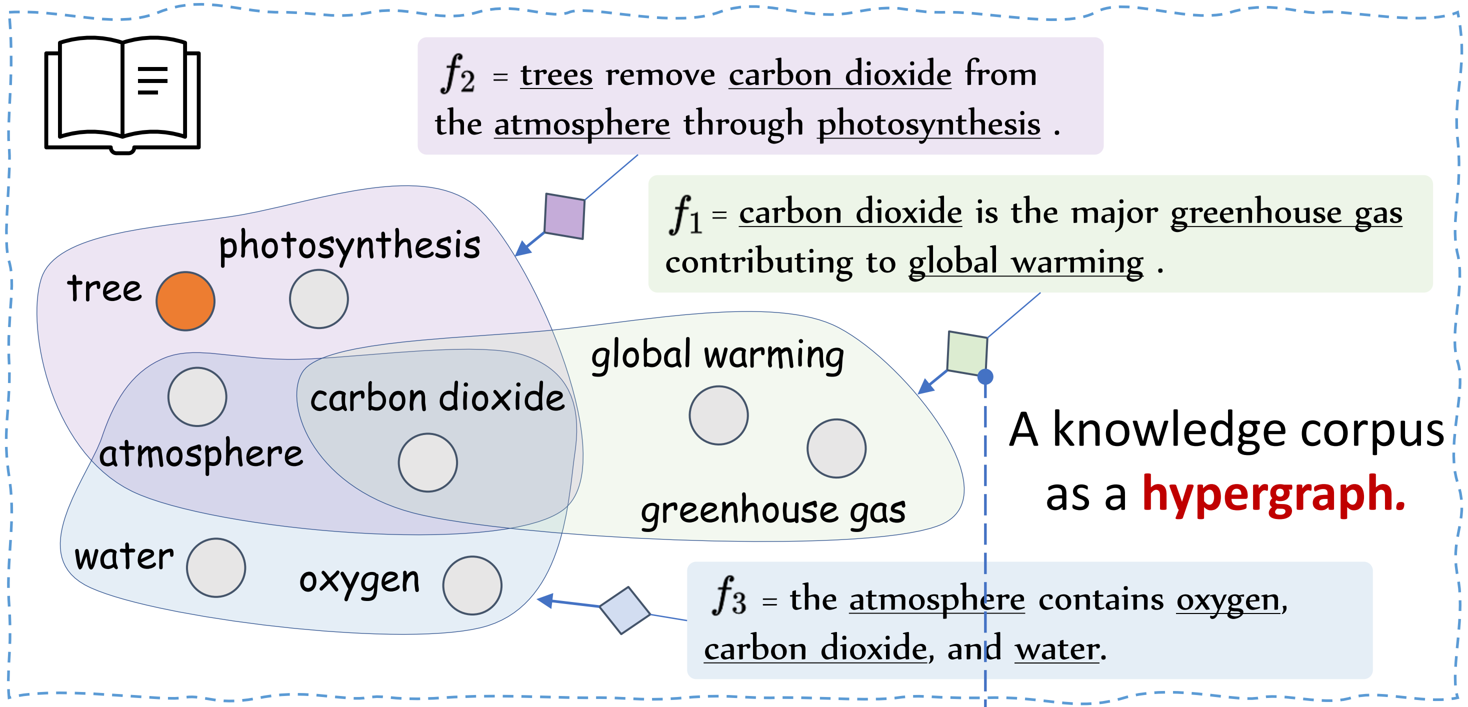
a **corpus** of common-sense facts, e.g., **GenericsKB**. $\mathcal{F}$

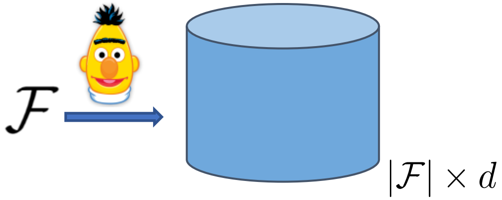
$f_i \in \mathcal{F}$

A **fact** is a sentence of generic commonsense knowledge

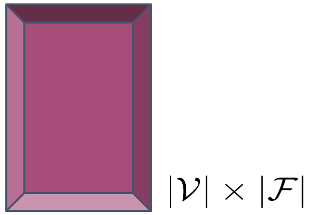
$c_j \in \mathcal{V}$

A **concept** is a noun or noun-chunk that are mentioned in $\mathcal{F}$

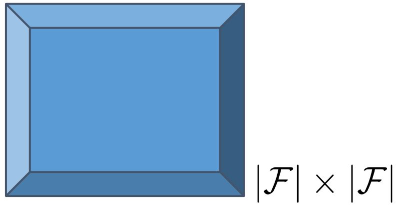




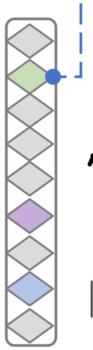
Dense Matrix of **Fact Embeddings**



Sparse Matrix of **Concept-to-Fact** Links



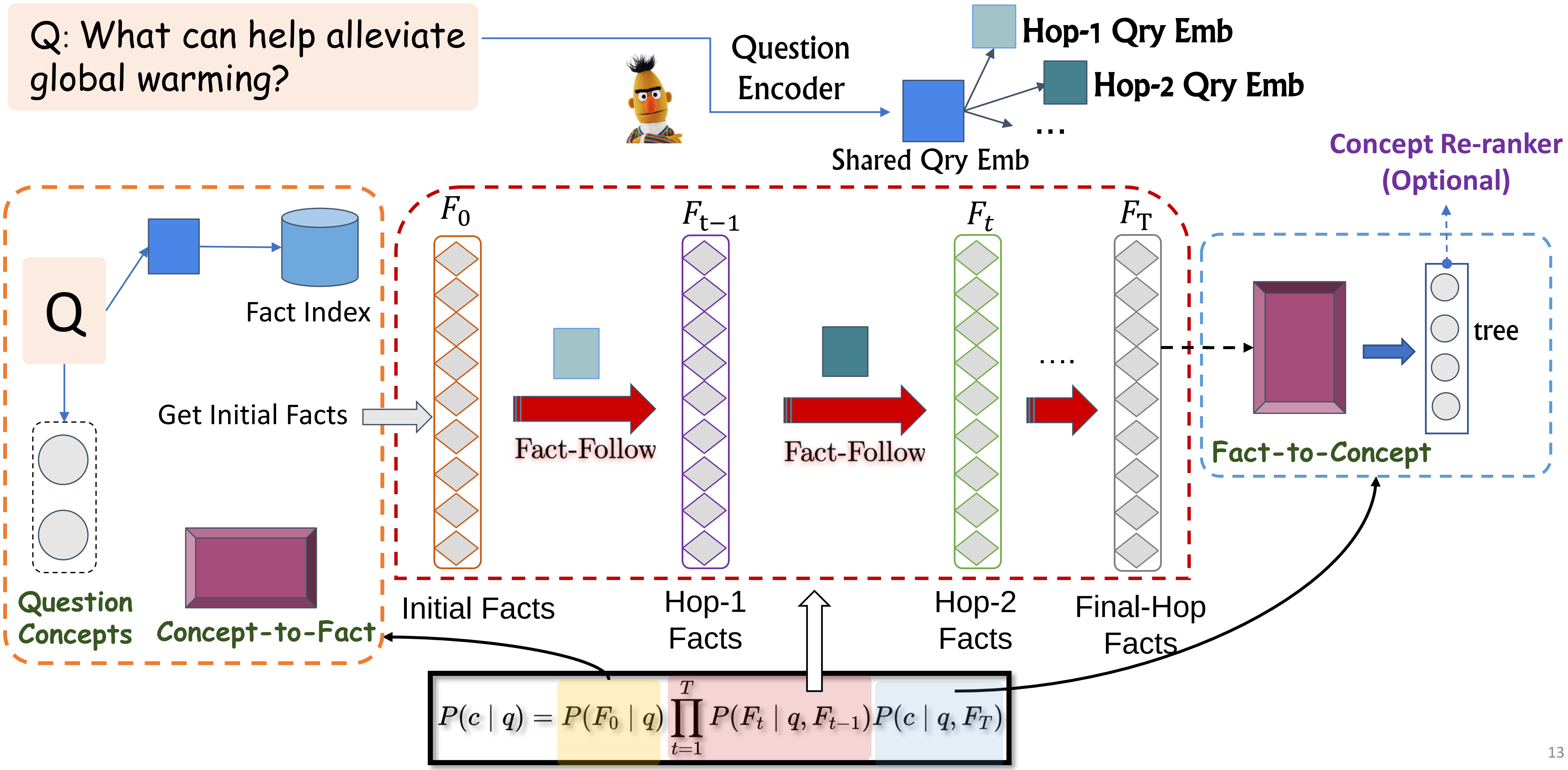
Sparse Matrix of **Fact-to-Fact** Links



A weighted set of facts → A sparse vector.

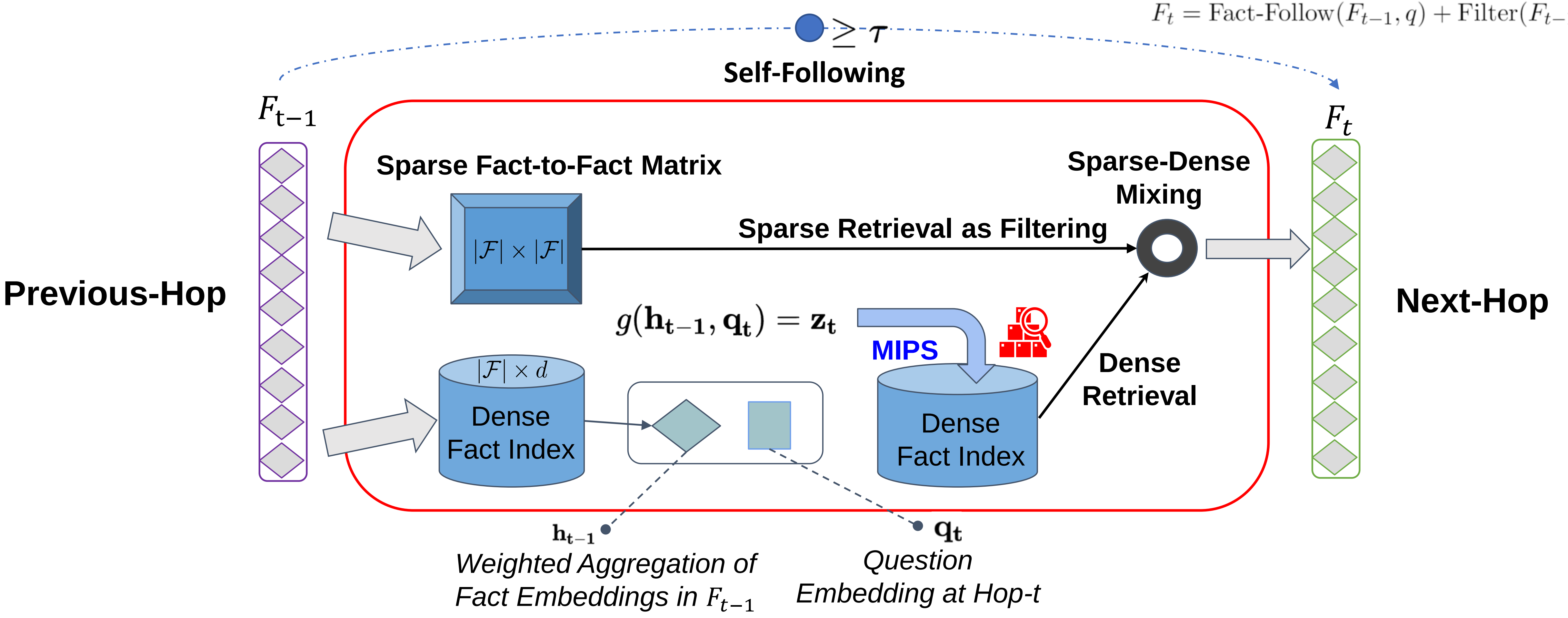
$|\mathcal{F}|$

Overall Workflow of DrFact



DrFact: Differentiable Fact-Following Operation

$$F_t = \text{Fact-Follow}(F_{t-1}, q) + \text{Filter}(F_{t-1}, \tau).$$



$$F_t = \text{Fact-Follow}(F_{t-1}, \mathbf{q_t})$$

Methods	BM25 (off-the-shelf)	DPR (EMNLP 2020)	DrKIT (ICLR 2020)	DrFact (NAACL 2021)
Knowledge Structure	A set of documents	A set of documents	Mention-Entity Bipartite Graph	Concept-Fact Hypergraph
Multi-hop Reasoning Formulation	N/A	N/A	Entity-Following	Fact-Following
Index for Dense Retrieval	N/A	Dense Fact Embeddings	Dense Mention Embedding	Dense Fact Embeddings
Sparse Retrieval Method	TF-IDF based Index+ BM25 Ranking Func.	N/A	Entity Cooccurrence	Fact-to-Fact Matrix
Multi-Hop Questions	N/A	N/A	Aggregating Multiple Models	A single model w/ Self-Following
Intermediate Supervision	N/A	N/A	N/A	Distant Supervision

Inference Efficiency:

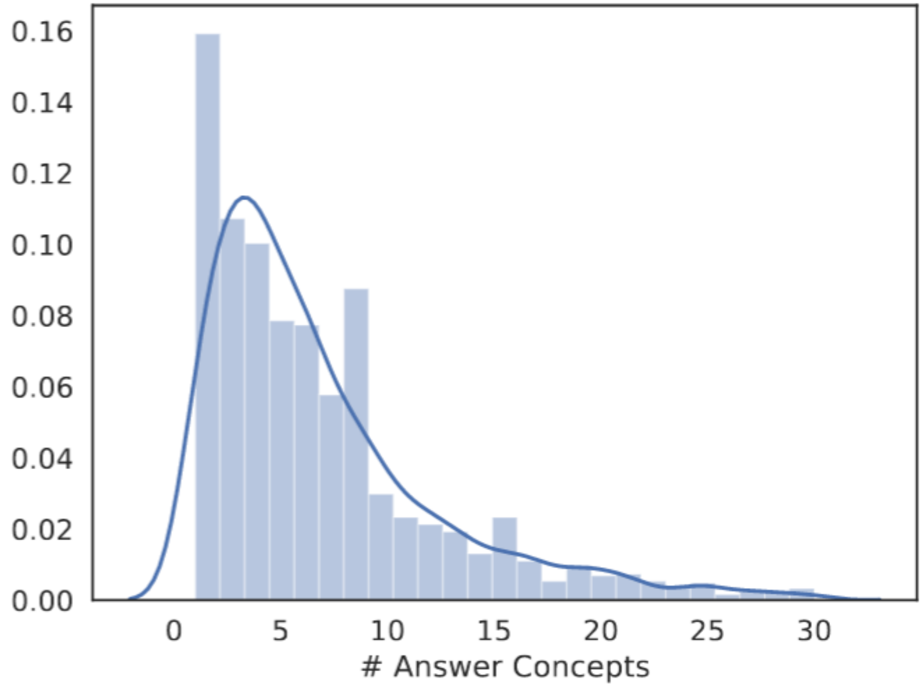
Methods	Major Computations	**Speed** (sec/q)
BM25	Sparse Retrieval	0.14
DPR	BERT-base + MIPS	0.08
DrKIT	BERT-base + T*(MIPS+ sp_{e2m})	0.47
DrFact	BERT-base + T*(MIPS+ sp_{f2f})	0.23
X+ MCQA	X + K * BERT-Large	+ *14.12*

T=3
K=300

Evaluation Setup

OpenCSR Dataset

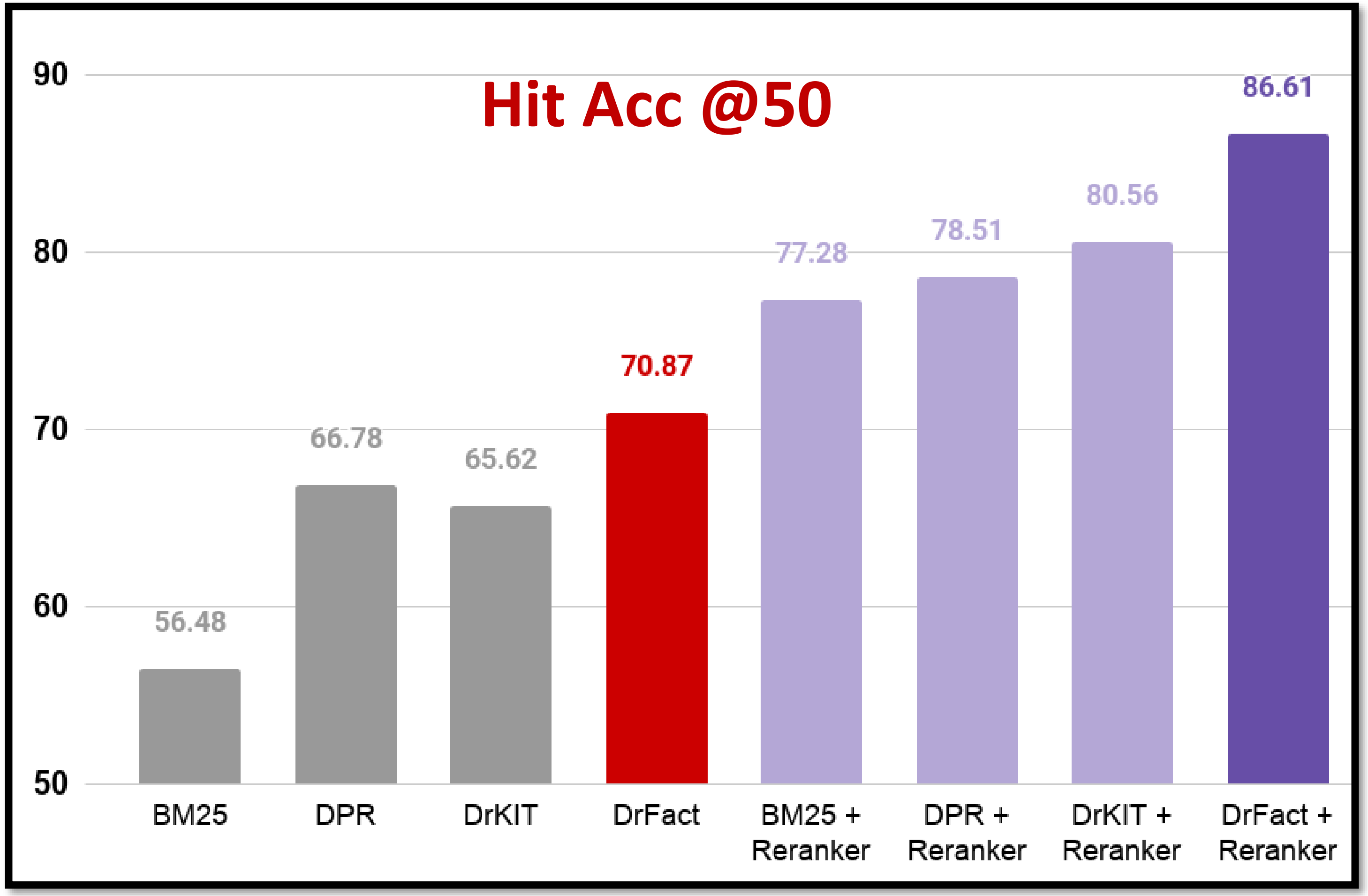
- Reformatted 3 Multiple-Choice QA Datasets (ARC, OBQA)
- + **Human-Annotated Answers** **(7 answer concepts per question on average)**

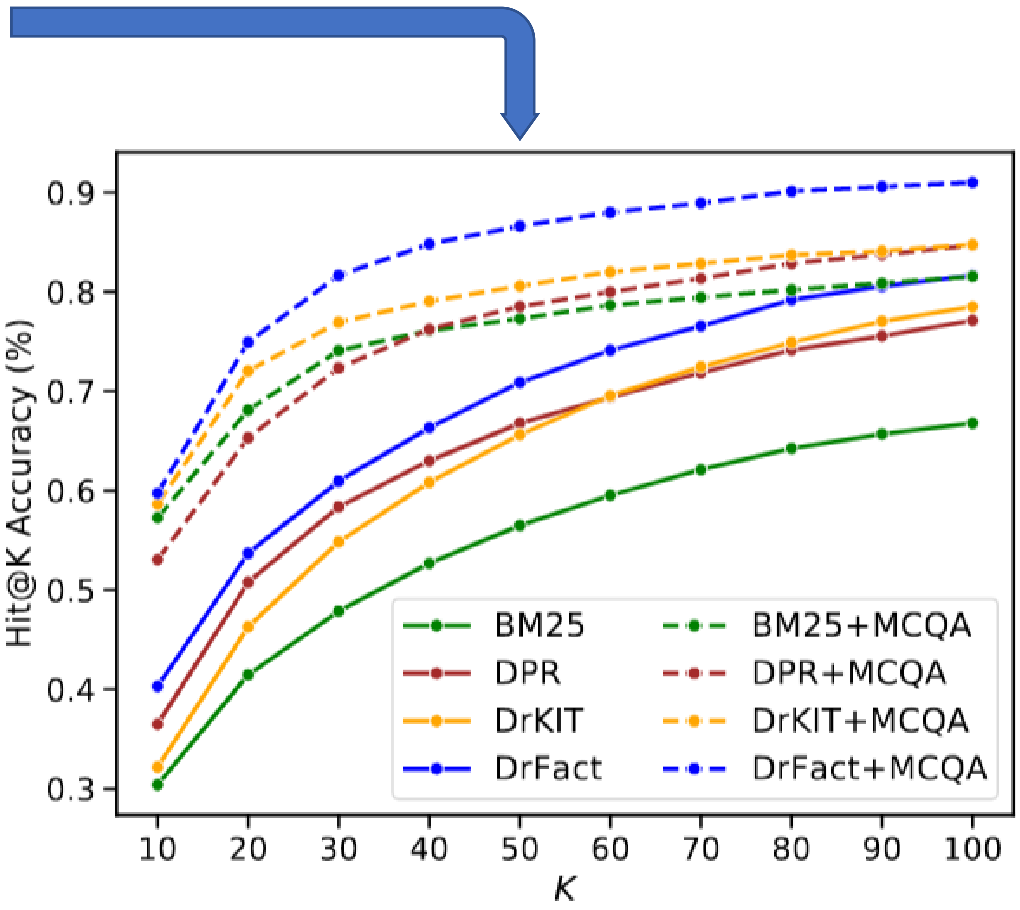


Metrics for evaluation

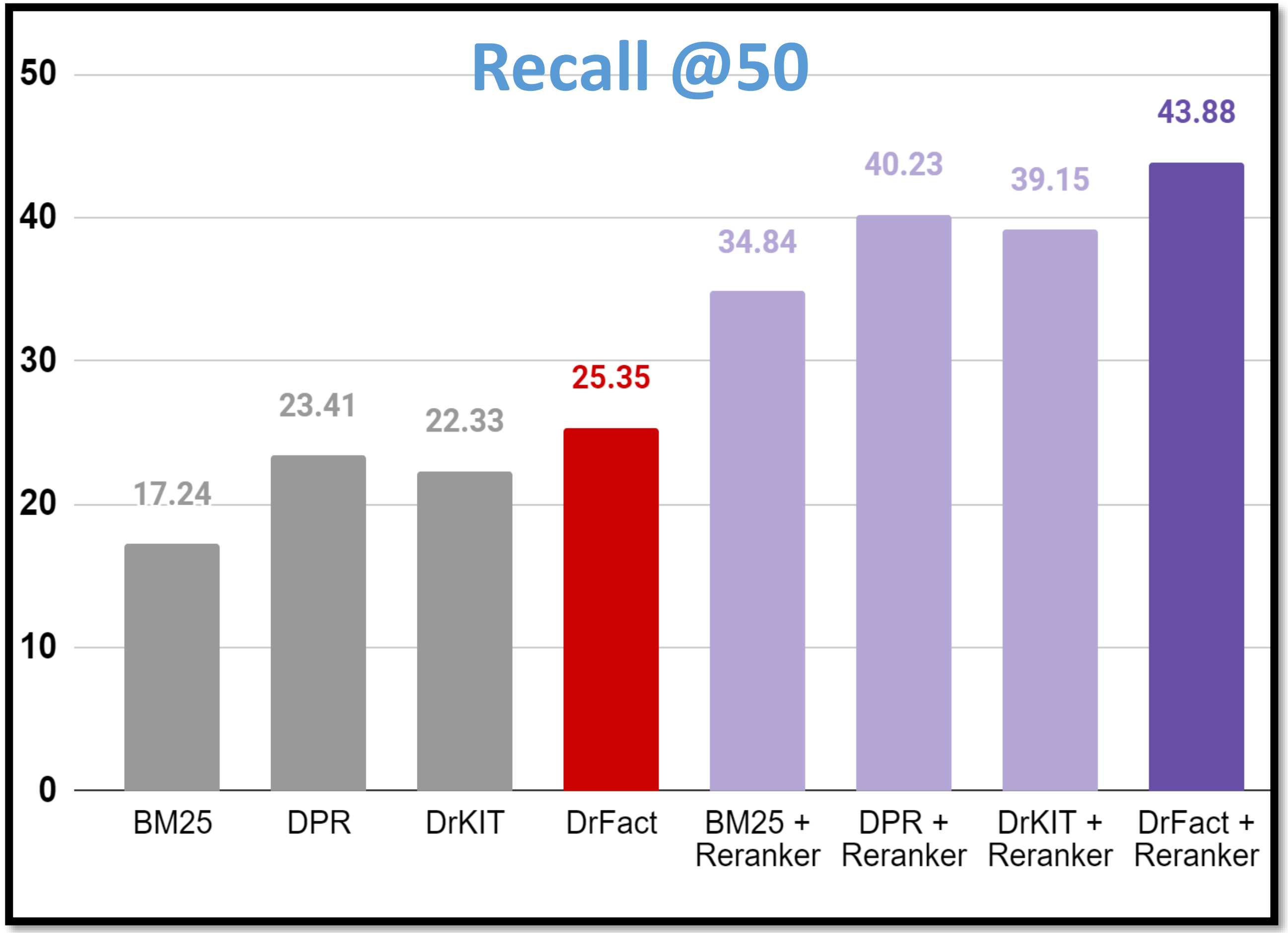
- **Hit Acc @ K <-- In the top K retrieved facts, there is at least one fact containing a correct answer (1 or 0).**
- **Ret Acc @ K <-- In the top K retrieved facts, the percentage of the covered answer concepts (over all the answer concepts).**
- Both are reported as an **average** over all examples in the **test set**.

Main Experimental Results





Experimental Results



DPR vs DrFact: Faithfulness and Interpretability

Q: "What will separate iron filings from sand? "

f1= angle irons reinforce the thinnest section of the ring ."

f2= sieves are used for separating fossils from sand..."

f3= stainless steel has a rough surface just after filing ." **DPR**

DrFact

iron filings show the *magnetic fields* . (in F0)

magnets *produce a magnetic field with a north ...* (in F1)

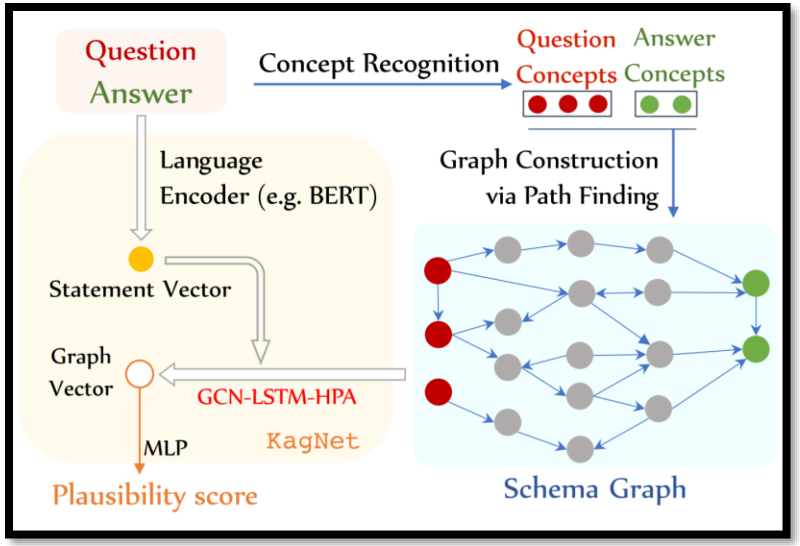
magnets *attract magnetic metals through magnetism* (in F2)

Findings and Take-Home Messages

- **OpenCSR is a novel setting to study CSR**
  - more **realistic** and **challenging** .
  - new **data** annotation for evaluation.

- **DrFact is an effective and efficient method for OpenCSR.**
  - **differentiable** Fact-Follow operation for **end-to-end** learning.
  - **state-of-the-art** performance comparing to strong baselines.
  - improve the explanations for multi-hop questions.

My PhD Progress on Common Sense Reasoning



KagNet & MHGRN: Knowledge-Aware Graph Networks for **Multiple-Choice CSR** (EMNLP 2019, 2020)

CommonGen for Generative CSR (EMNLP Findings 2020)

This talk. (NAACL-HLT 2021).

Novel Benchmarks

Future

Closed-Ended CSR Methods

Open-Ended CSR

NumerSense (EMNLP 2020)
Probing Numerical Commonsense Knowledge from BERT

RiddleSense (under review, on arXiv, 2021)
Testing Machines for Solving Puzzling Riddles

Future Directions

- Common-Sense Reasoning **Beyond English**
  - Common sense knowledge as the bridge for breaking language barriers.

- **Embodied Intelligence** w/ Common Sense
  - Learning to make sequential actions in interactive physical environment.

- **Open-Source Toolkit** for Common Sense Reasoning
  - Connecting commonsense research with realistic application scenarios.