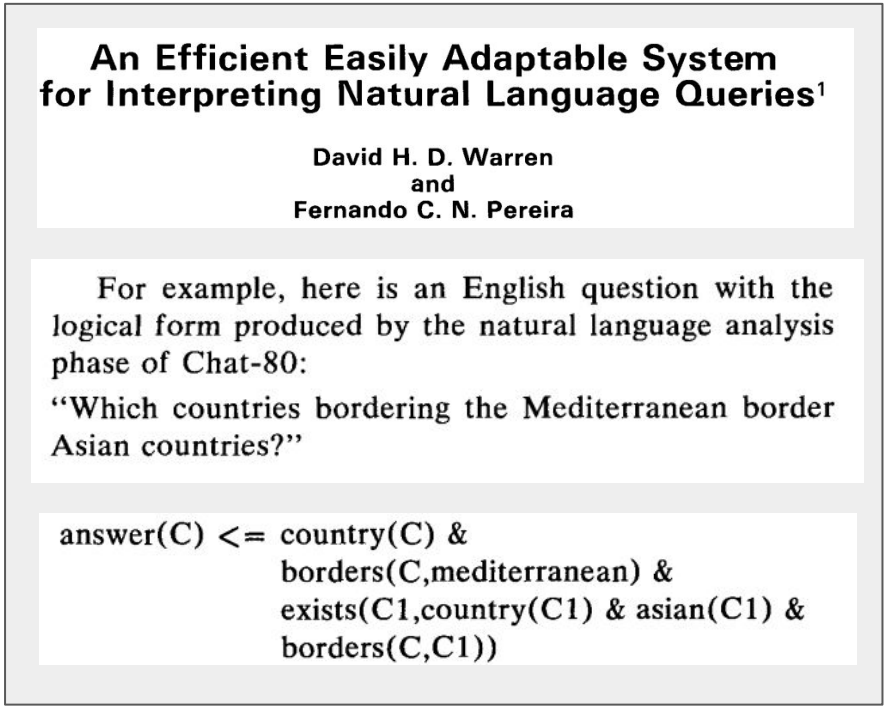
Google DeepMind

Extracting Structure From Text as Questions, Entities, and Answers

William W. Cohen
with Wenhu Chen, Pat Verga, Haitian Sun, Nitish Gupta, Michiel De Jong, John Wieting, and others

Motivation and Background

Question-Answering with a KB: A Timeline of KBQA



An Efficient Easily Adaptable System for Interpreting Natural Language Queries[1]

David H. D. Warren
and
Fernando C. N. Pereira

For example, here is an English question with the logical form produced by the natural language analysis phase of Chat-80:

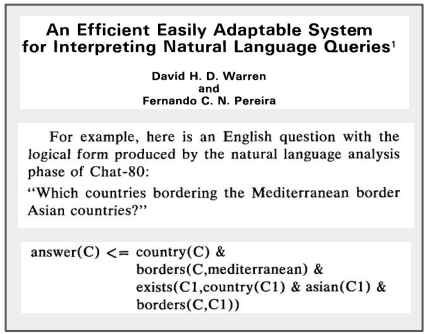
"Which countries bordering the Mediterranean border Asian countries?"

```
answer(C) <= country(C) &
             borders(C,mediterranean) &
             exists(C1,country(C1) & asian(C1) &
             borders(C,C1))
```

Chat-80
Warren &
Pereira

1980

KBQA: A Timeline



An Efficient Easily Adaptable System for Interpreting Natural Language Queries[1]

David H. D. Warren and Fernando C. N. Pereira

For example, here is an English question with the logical form produced by the natural language analysis phase of Chat-80:

"Which countries bordering the Mediterranean border Asian countries?"

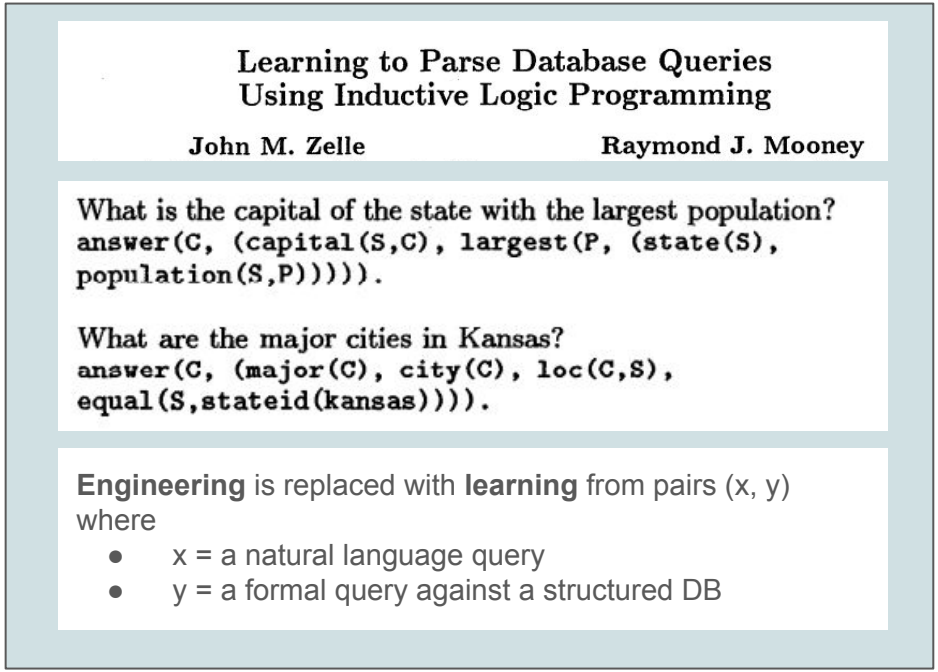
```
answer(C) <= country(C) &
             borders(C,mediterranean) &
             exists(C1,country(C1) & asian(C1) &
             borders(C,C1))
```

Chat-80
Warren & Pereira

1980

CHILL
Zelle & Mooney

1996



Learning to Parse Database Queries Using Inductive Logic Programming

John M. Zelle Raymond J. Mooney

What is the capital of the state with the largest population?

```
answer(C, (capital(S,C), largest(P, (state(S),
population(S,P))))).
```

What are the major cities in Kansas?

```
answer(C, (major(C), city(C), loc(C,S),
equal(S,stateid(kansas)))).
```

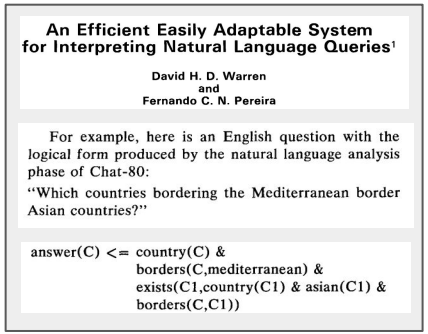
Engineering is replaced with **learning** from pairs (x, y) where

- x = a natural language query
- y = a formal query against a structured DB

"Which school that Sir Ernest Rutherford attended has the latest founding date?"
"Which of the countries bordering Mexico have an army size of less than 1050?"
"Where is the end of the river that originates in Shannon Pot?"

Google DeepMind

KBQA: A Timeline



An Efficient Easily Adaptable System for Interpreting Natural Language Queries[1]

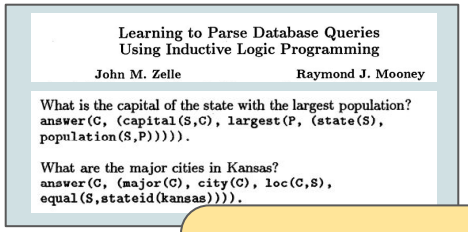
David H. D. Warren
and
Fernando C. N. Pereira

For example, here is an English question with the logical form produced by the natural language analysis phase of Chat-80:

"Which countries bordering the Mediterranean border Asian countries?"

```
answer(C) <= country(C) &
             borders(C,mediterranean) &
             exists(C1,country(C1) & asian(C1) &
             borders(C,C1))
```

Chat-80
Warren & Pereira



Learning to Parse Database Queries Using Inductive Logic Programming

John M. Zelle Raymond J. Mooney

What is the capital of the state with the largest population?

```
answer(C, (capital(S,C), largest(P, (state(S),
population(S,P))))).
```

What are the major cities in Kansas?

```
answer(C, (major(C), city(C), loc(C,S),
equal(S,stateid(kansas)))).
```

KBQA for FreeBase, a large *open-domain* KB that integrated many other smaller KBs

Questions that exploit the *modularity* of KBs: assertions in a KB can be *combined*

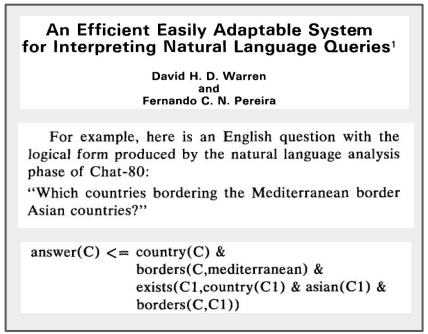
Complex WebQuestions
Talmor & Berant

KBQA with complex logical questions about virtually any information stored in the world's largest KB!

Modularity implies that the number of answerable questions is **exponentially more** than the number of stored facts

1980 1996 2013 2018

KBQA: A Timeline



An Efficient Easily Adaptable System for Interpreting Natural Language Queries[1]

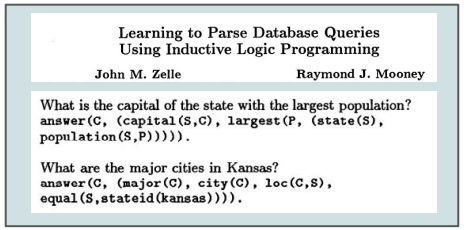
David H. D. Warren and Fernando C. N. Pereira

For example, here is an English question with the logical form produced by the natural language analysis phase of Chat-80:

"Which countries bordering the Mediterranean border Asian countries?"

```
answer(C) <= country(C) &
             borders(C,mediterranean) &
             exists(C1,country(C1) & asian(C1) &
             borders(C,C1))
```

Chat-80
Warren & Pereira



Learning to Parse Database Queries Using Inductive Logic Programming

John M. Zelle Raymond J. Mooney

What is the capital of the state with the largest population?

```
answer(C, (capital(S,C), largest(P, (state(S),
population(S,P))))).
```

What are the major cities in Kansas?

```
answer(C, (major(C), city(C), loc(C,S),
equal(S,stateid(kansas)))).
```

CHILL
Zelle & Mooney

WebQuestions
Berant et al

Complex WebQuestions
Talmor & Berant

Natural Questions
Kwaitkowski et al

1980 1996 2013 2018 19

Google DeepMind

"Natural" = relatively complex Google queries

KBQA: A Timeline

For example, here is an English question with the logical form produced by the natural language analysis phase of Chat-80:

"Which countries bordering the Mediterranean border Asian countries?"

```
answer(C) <= country(C) &
             borders(C,mediterranean) &
             exists(C1,country(C1) & asian(C1) &
             borders(C,C1))
```

Chat-80
Warren & Pereira

CHILL
Zelle & Mooney

Many of the questions people **actually ask** are **not answerable** using **existing KBs.**

Natural Questions
Kwaitkowski et al

when are hops added to the brewing process?

what does the word china mean in chinese?

where do dust storms occur in the US?

where is blood pumped after it leaves the right ventricle?

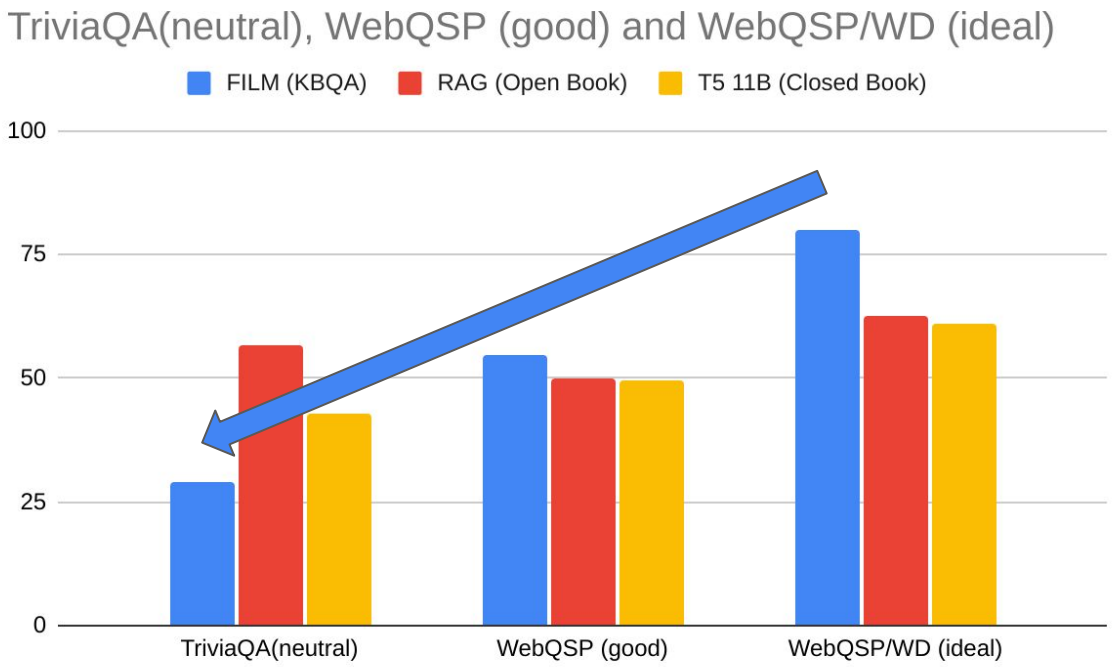
1980 1996 2013 2018

KBQA only works for the “right” questions

Four datasets, three methods:

- Natural Questions
- Trivia QA: mostly entity-related trivia questions.
- WebQuestionsSP: must be FreeBase answerable and have a semantic parse
- WebQuestionsSP/WD - also have answers in a 1M-entity WikiData subset

Methods: **FILM** uses KB; **RAG** retrieves *documents* and uses NLP to read them; **T5 11B** answers using model parameters

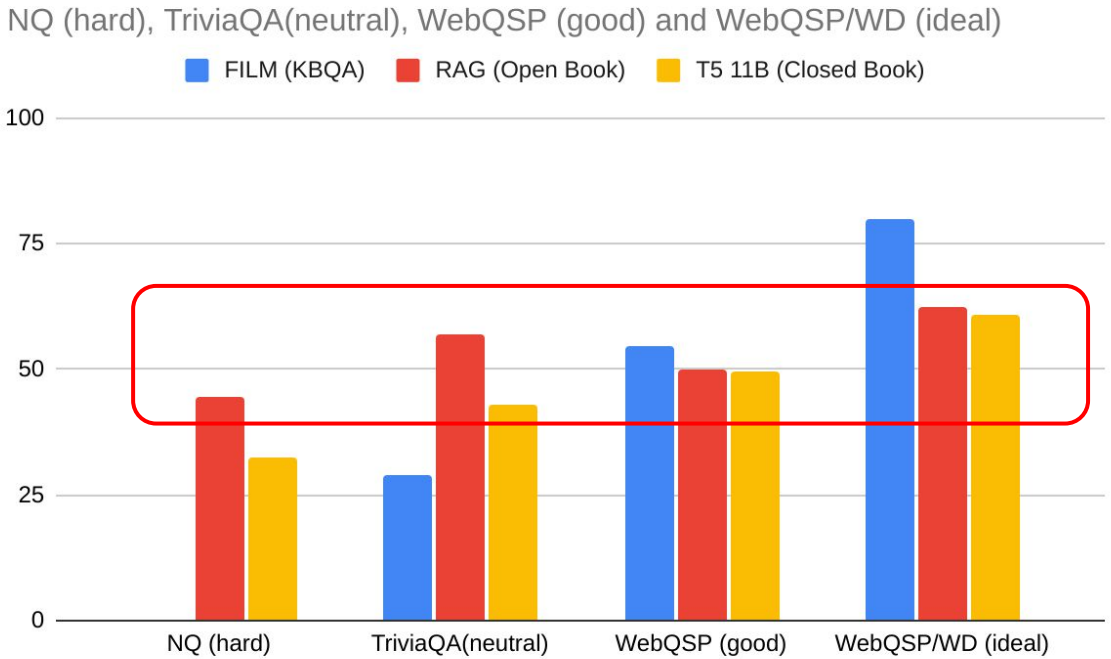


Text-based “Open Book” QA is more robust

Four datasets, three methods:

- Natural Questions
- Trivia QA: mostly entity-related trivia questions.
- WebQuestionsSP: must be FreeBase answerable and have a semantic parse
- WebQuestionsSP/WD - also have answers in a 1M-entity WikiData subset

Methods: **FILM** uses KB; **RAG** retrieves *documents* and uses NLP to read them; **T5 11B** answers using model parameters



Is open book QA >> KBQA?

No: Open-book QA still has trouble with compositionality

Complex WebQuestions - with KB

Model	Accuracy
CBR-KBQA	70.4
NSM+h	53.9
DynAS	50
PullNet	45.9
QGG (Query Graph Generator)	44.1
KBQA-GST	39.4

Complex WebQuestions - with Text

Model	Precision
Human	63
*SplitQA + data augmentation	34.2
MHQA-GRN	30.1
SPLITQA + PRETRAINED	25.9
SIMPQA + PRETRAINED	19.9

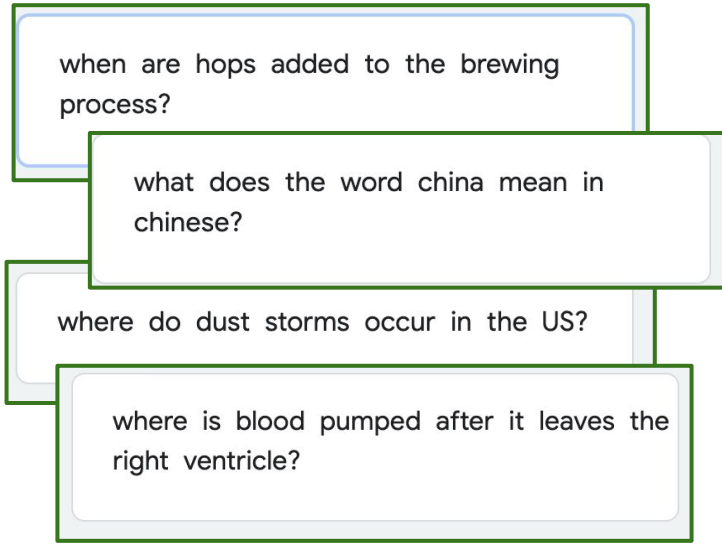
Best of both worlds?

"Which school that Sir Ernest Rutherford attended has the latest founding date?"
"Which of the countries bordering Mexico have an army size of less than 1050?"
"Where is the end of the river that originates in Shannon Pot?"

Can we build KR/QA systems that

- exploit the modularity of KBs to store and query information **compositionally and**
- have good coverage on the questions that people **naturally** ask?

Our proposal: a KB of **question-answer pairs** generated **automatically** from text.

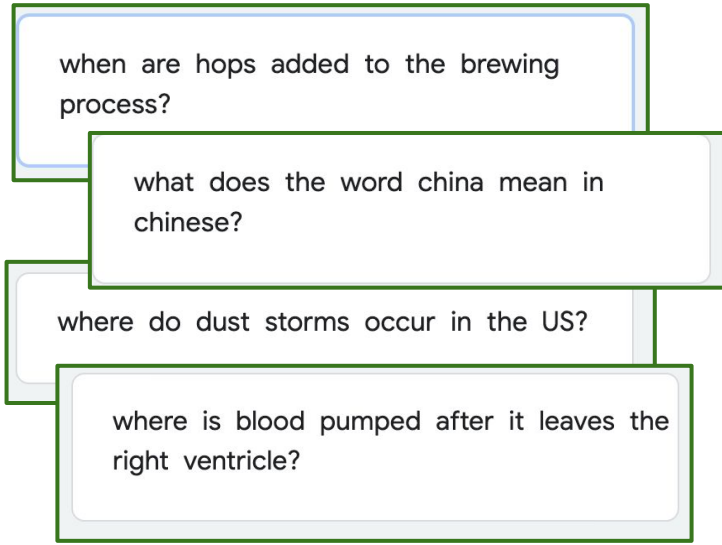


Best of both worlds?

"Which school that Sir Ernest Rutherford attended has the latest founding date?"
"Which of the countries bordering Mexico have an army size of less than 1050?"
"Where is the end of the river that originates in Shannon Pot?"

Our proposal: a KB of **question-answer pairs** generated **automatically** from text.

- KB extraction and QA system (Chen et al AAAI 2023, Chen et al EACL 2023)
- Ongoing work on KB extraction (Sun et al, 2023)



Questions, Answers, and Entities as a KB

Generating questions and answers from a KB

QA datasets (like SQuaD and NQ) contain many triples (question, passage, answer), so you can train a model Pr(answer | question, passage)

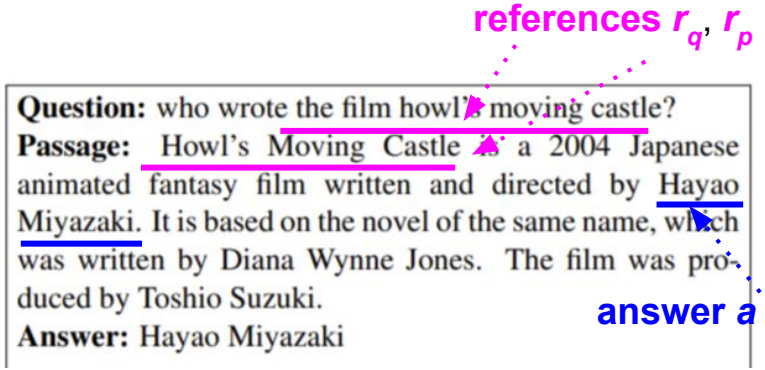
Alternatively you can train a model Pr(question, answer | passage) and generate many Q-A pairs from a passage.

Question: who wrote the film howl's moving castle?
Passage: Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki. It is based on the novel of the same name, which was written by Diana Wynne Jones. The film was produced by Toshio Suzuki.
Answer: Hayao Miyazaki

Generating questions and answers from a KB

Additional structure has been added to these annotated corpora to support **explainable question answering** [Lamm, et al, "QED" 2021].

- Question and answer contain annotated "references" (often entities)
- Equality between references is annotated.
  - Entity links from the passage can be propagated to the questions.



The annotations (r_p, q^*, a) are like a KB triple:

(Howls_Moving_Castle$_{Q42713}$, film.writer, Hyao_Miyazaki$_{Q743313}$)

Abstract version of question q*: who wrote $1?

QEDB: Key Ideas (1) Annotated questions ~= KB relation

Peter Dinklage$_1$
.... is an American actor and producer. He received acclaim for portraying Tyrion Lannister on the television series Game of Thrones (2011–2019)...
1

Who is the actor who plays tyrion in game of thrones

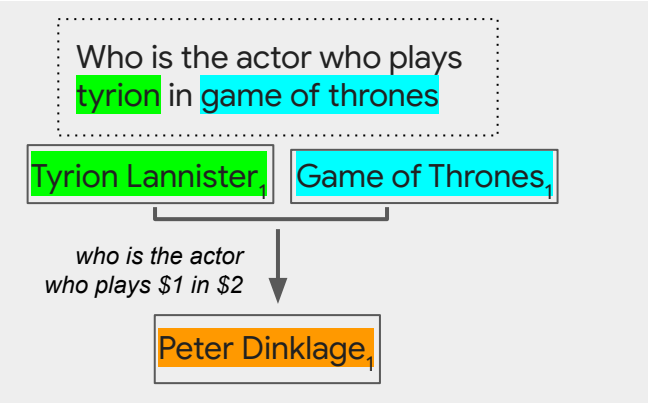
↓

Peter Dinklage$_1$

- A question-answer pair (with **annotations**) describe a **relationship** between NPs in the answer passage (doc 1)
  - Peter Dinklage$_1$, Tyrion Lannister$_1$, and Game of Thrones$_1$ are **related**

QEDB: Key Ideas (1) Annotated questions ~= KB relation

Peter Dinklage$_1$
.... is an American actor and producer. He received acclaim for portraying Tyrion Lannister on the television series Game of Thrones (2011–2019)...
1

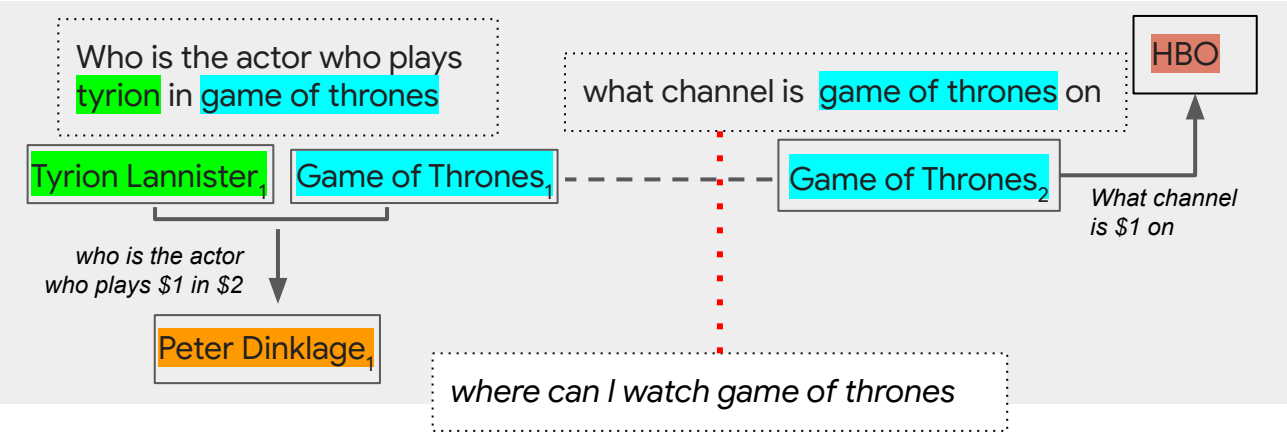


- A question-answer pair (with **annotations**) describe a **relationship** between NPs in the answer passage (doc 1)
  - Peter Dinklage$_1$, Tyrion Lannister$_1$, and Game of Thrones$_1$ are **related**
- **Question explanations** are **modular, atomic** objects - a lot like **KG relations!** But each QA explanation is **aligned with user questions**
  - Information is written about and also *asked about*

QEDB: Key Ideas (2) Large-scale QG + crosslinking creates a "KB"

Peter Dinklage$_1$
.... is an American actor and producer. He received acclaim for portraying Tyrion Lannister on the television series Game of Thrones (2011–2019)...
1

Game of Thrones
Game of Thrones is an American fantasy drama television series created by David Benioff and D. B. Weiss for HBO.
2



Currently using

1. **Entity linking** is on passages in Wikipedia
   - Lots of context!
2. About ½ of question phrases **are** entities and most **contain** entities
3. **60M** generated, filtered Wikipedia questions from the Facebook PAQ project**
4. T5-based question annotators trained on QED/NQ data

We also can soft-match questions.

QEDB: A KB of grounded, atomic relationships based on *what is asked*

Peter Dinklage$_1$
.... is an American actor and producer. He received acclaim for portraying Tyrion Lannister on the television series Game of Thrones (2011–2019)... his breakthrough came as Finbar Mcbride in the comedy-drama The Station Agent (2003)
1

Game of Thrones
Game of Thrones is an American fantasy drama television series created by David Benioff and D. B. Weiss for HBO. It is an adaptation of a series of fantasy novels by George R. R. Martin.
2

Rebels, Rogues & Sworn Brothers
...The name of the album comes from the appendix of Storm of Swords , the third book in George R.R. Martin's Song of Ice and Fire series.
3

Who is the actor who plays tyrion in game of thrones

Tyrion Lannister$_1$ | Game of Thrones$_1$

who is the actor who plays $1 in $2

Peter Dinklage$_1$

who played $1 in $2

Finbar McBride | The comedy-drama The Station Agent (2003)

what channel is game of thrones on

HBO

Game of Thrones$_2$

What channel is $1 on

what is game of thrones based on

what is $1 based on

a series of fantasy novels by George R. R. Martin

Storm of Swords$_3$

what is the third book in george r r martin's a song of ice and fire

What is the third book in $1 's $2

George R. R. Martin$_3$ | A Song of Ice and Fire$_3$

who is the author of $1

who is the author of a song of ice and fire

who wrote a song of ice and fire

Claim 1: Components in the QEDB are accurate

Query Entity	Related Entity	Linking Question
Tonga	Deborah Ann Gardner	*in which country was deborah gardner murdered by the peace corps in 1976*
	1976	*in which country was deborah gardner murdered by the peace corps in 1976*
	Fiji	*where did the sulu come from in fiji*
	American Samoa	*what country did american samoa defeat to win their first ever world cup*
	2011	*what country did american samoa lose to in 2011*
	New Zealand	*in 2006, australian troops and new zealand police were sent to which island*
	Tropical cyclone	*which island nation was hardest hit by cyclone gita*
	1969	*which country issued a banana shaped postage stamp from 1969 to 1985*
	Samoa	*who did samoa play in the under 20 world cup*
	Niuatoputapu	*niuatoputapu is the highest point in which pacific island nation*
arXiv	Preprint	*where can you find a preprint repository*
	Ricci flow and the Poinc (GPC)	*on which computer system is ricci flow and the poincaré conjecture*
	An Exceptionally Simple …	*where is an exceptionally simple theory of everything located*
	VAN method	*where is the van method of measuring earthquakes stored*
	ARC fusion reactor	*where did the idea of the arc fusion reactor come from*
	Database	*what is an example of a preprint database*
	Mathematics	*who invented the first open access journal in the history of mathematics*
	Particle	*where was the article about quarkonium published*
	Results in Mathematics	*on what platform have some important results in mathematics been published*
	Particle physics	*where can preprints be found in physics*
Jerry Garcia	Grateful Dead	*which member of the grateful dead died in 1995*
	Guitar	*who played guitar on three tracks by ornette coleman in 1988*
	Lead guitar	*who was the lead guitarist for the dead in europe in 1972*
	Pedal steel guitar	*who plays guitar on teach your children by graham nash*
	~~Today (Song)~~	*who plays the lead guitar on today by jefferson airplane*
	*Roseanne (TV show character)	*what's the name of roseanne's fourth child* *answer is actually Jerry Garcia Conner
	~~Teach Your Children~~	*who plays guitar on teach your children by graham nash*
	1995	*which member of the grateful dead died in 1995*
	The Grateful Dead Movie	*who directed the movie the grateful dead movie*
	Sandy Rothman	*sandy rothman was a close friend and collaborator of which grateful dead guitarist*

Claim 2: Components in the QEDB are modular

QEDB

Question 1	Bridge Entity	Question 2	Answer
where did the cincinnati reds last game	Joe's North ...	*what nfl team used to play at $1*	Cincinnati Bengals
who sings the song please leave the grates	Jebediah	*when was $1 formed and by whom*	1994
what event was hosted in 2012 by marseille	World Water Forum	*who organizes $1*	World Water Council
who is the main character of the tombs of atuan	Tenar	*who raised $1 in wizard of earthsea lore*	Aihal
*what is the sequel to wild fire by nelson de mille**	Night Fall	*when did the plane crash in $1*	1996
who directed the opening act	Steve Byrne	*$1 is the lead actor in which us tv series*	Sullivan & Son
who designed the set for le piege de meduse	Willem de Kooning	*$1 is an example of what*	abstract expressionism
who maintains the lights on the isle of man	N. Lighthouse Boar	*$1 is in which country*	Scotland
who was the roman proponent of hedonism	Lucretius	*what is the name of $1's book on atomism*	On the Nature of Things
what is the main export of tutuila	Canned fish	*what is ... the process used to preserve $1*	Canning
who was the creator of "the rocky horror show"	Richard O'Brien	*when was $1 born*	1942
which nickelodeon show was bill long credited in	Blues Clues	*who originally hosted $1*	Steve Burns
what college was thomas balaton educated at after eton	New College	*when was $1 founded*	1379
which actor composed the song "smile" from 1st round	Charlie Chaplin	*what nationality is $1*	English
what casino is formerly known as vegas world	Stratosphere Las Vegas	*what company is headquartered at $1*	Amer. Casino & Entertain...

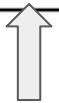
HotpotQA

Claim 2: Components in the QEDB are modular

…and we can answer multi-hop questions with these components!

QAMAT and QAMAT+ [Chen, et al EACL 2023 https://arxiv.org/abs/2204.04581]

Model (Dev Set F1 Score)	HoPoQ	MusQ
T5-3B (Roberts et al., 2020)	27.8	7.5
T5-11B (Roberts et al., 2020)	30.2	9.0
MDR+T5-Decoder (Xiong et al., 2020)	62.6	26.8
RePAQ (Lewis et al., 2021)†	47.8	18.6
QAMAT	42.0	16.7
QAMAT+	57.6	29.8



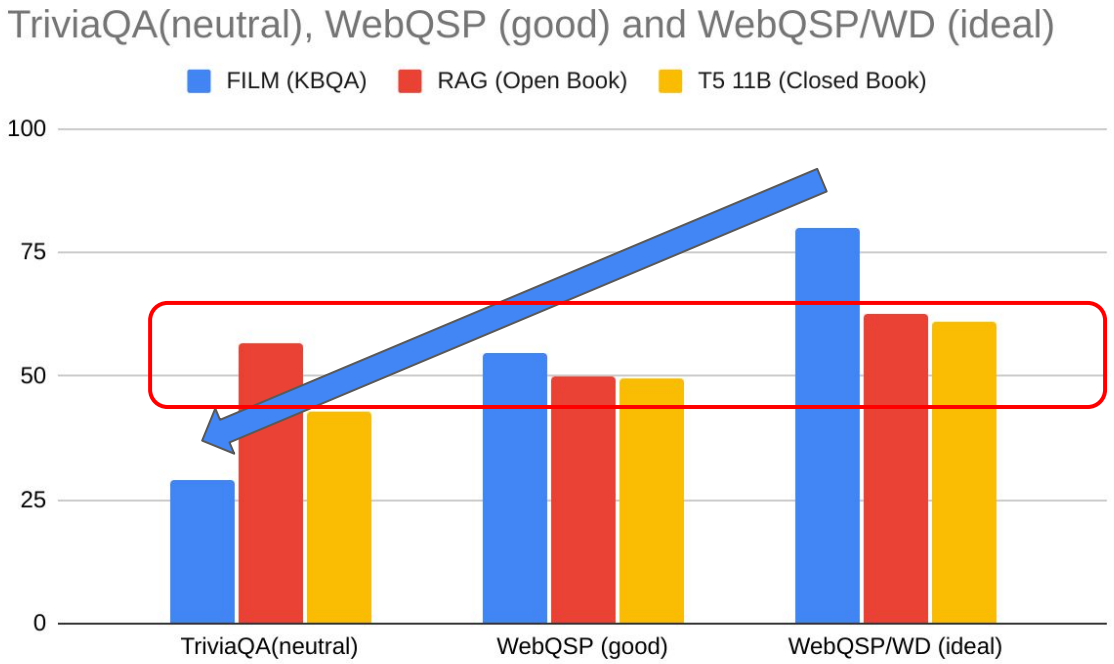
who was the creator of "the rocky horror show"	Richard O'Brien	*when was $1 born*	1942
which nickelodeon show was bill long credited in	Blues Clues	*who originally hosted $1*	Steve Burns
what college was thomas balaton educated at after eton	New College	*when was $1 founded*	1379
which actor composed the song "smile" from 1st round	Charlie Chaplin	*what nationality is $1*	English
what casino is formerly known as vegas world	Stratosphere Las Vegas	*what company is headquartered at $1*	Amer. Casino & Entertain…

**Prior work [Lewis et al, 2021, PAQ dataset and RePAQ system] explored answering questions with a DB of question-answer pairs, but not the possibility of linking QA pairs into a modular KB.

Claim 3: KBQA using a QEDB is aligned with users' questions

Recap: KBQA only works for the "right" questions

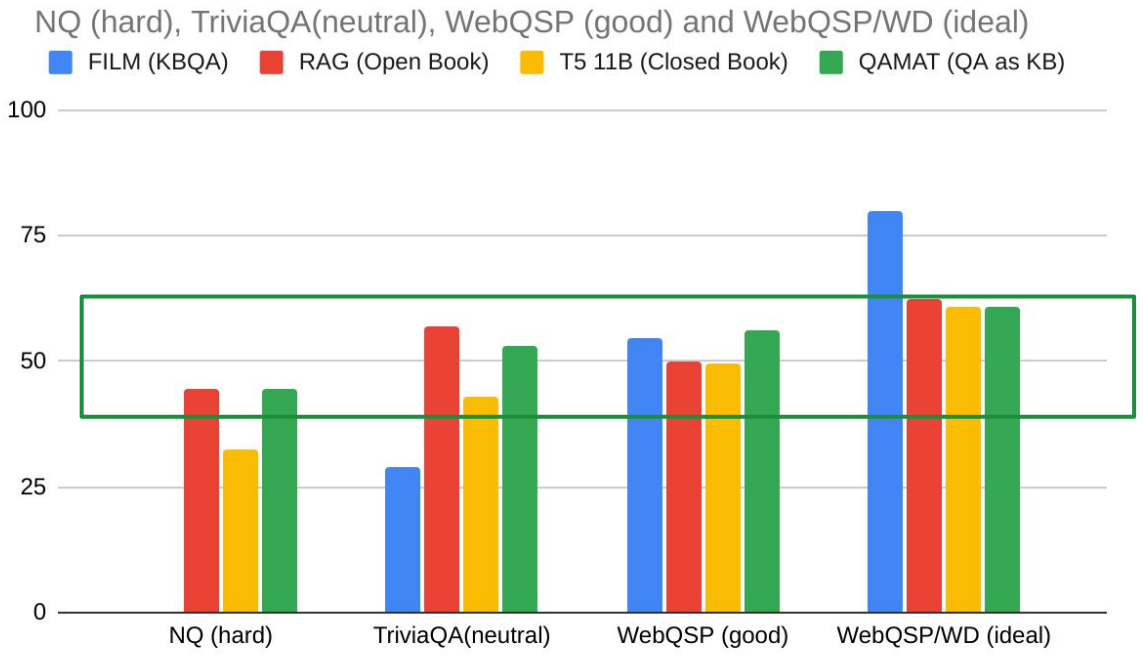
- Trivia QA: mostly entity-related trivia questions.
- WebQuestionsSP: must be FreeBase answerable and have a semantic parse
- WebQuestionsSP/WD - also have answers in a 1M-entity WikiData subset



Claim 3: KBQA using a QEDB is aligned with users' questions

Recap: KBQA only works for the "right" questions

- NQ: natural questions
- Trivia QA: mostly entity-related trivia questions.
- WebQuestionsSP: must be FreeBase answerable and have a semantic parse
- WebQuestionsSP/WD - also have answers in a 1M-entity WikiData subset



Claim 3: The QEDB is aligned with users' questions

Augmenting Pre-trained Language Models with QA-Memory Chen, et al EACL 2023
https://arxiv.org/abs/2204.04581

Performs well vs comparably sized memory-augmented models, or larger models using closed-book approaches

Outperforms prior work using QA memories (RePAQ)

	hard	neutral	good
Model (Test Set)	NQ	TQA	WQ
T5-3B (Roberts et al., 2020)	30.4	35.1	33.6
T5-11B (Roberts et al., 2020)	32.6	42.3	37.2
EaE (Févry et al., 2020)	-	43.2	-
FilM (Verga et al., 2021)	-	29.1	-
TOME-2 (de Jong et al., 2022)	-	53.4	-
DensePhrases (Lee et al., 2021)	40.9	50.7	-
REALM (Guu et al., 2020)	40.4	55.8	40.7
DPR (Karpukhin et al., 2020)	41.5	57.9	42.4
RAG-Seq (Lewis et al., 2020)	44.5	56.8	45.2
FiD (Izacard and Grave, 2021)	48.2	65.0	-
RePAQ (Lewis et al., 2021)	41.2	38.8	29.4†
QAMA Fine-tuned (Base)	44.5	53.2	43.0
QAMA Fine-tuned (Large)	45.5	54.8	43.6

Aside: this problem is *not* solved by recent LLMs

NQ leaderboard

LLaMA 65B (few-shot, k=64)	39.9
PaLM-540B (Few-Shot, k=64)	39.6
PaLM 2-L (one-shot)	37.5

Chinchilla (few-shot, k=64)	35.5
LLaMA 65B (few-shot, k=5)	35.0
GLaM 62B/64E (Few-Shot)	32.5

PaLM 2-M (one-shot)	32.0
LLaMA 65B (one-shot)	31.0
GPT-3 175B (Few-Shot, k=64)	29.9

Claim 3: The QEDB is aligned with users' questions

An **Efficient** Memory-Augmented Transformer for Knowledge-Intensive NLP Tasks:
Wu, Zhao, Hu, Minervini, Stenetorp, Riedel, 2022

Long-form dialog

Model	F1	R-L	U/s
Parametric models			
Trans MemNet (Dinan et al., 2019)	11.85	10.11	-
BART-large (Lewis et al., 2020a)	12.86	11.77	55
T5-base (Raffel et al., 2020)	13.53	12.40	160
Retrieval-augmented models			
BART + DPR (Petroni et al., 2021)	15.19	13.23	0.7
RAG (Lewis et al., 2020b)	13.11	11.57	3.4
Retrieval-only models			
RePAQ w/ EMAT key encoder	1.84	1.48	-
Ours			
EMAT-FKSV	**15.78**	**14.73**	141
EMAT-SKSV	15.35	14.68	**150**

Table 2: Results on the Wizard-of-Wikipedia dataset from the KILT benchmark.

Model	NQ EM	NQ Q/s	TQA EM	WQ EM
Parametric models				
T5-base (Roberts et al., 2020)	25.8	1600	24.4	26.6
T5-large (Roberts et al., 2020)	27.6	570	29.5	27.7
T5-3B (Roberts et al., 2020)	30.4	55	35.1	33.6
T5-11B (Roberts et al., 2020)	32.6	-	42.3	37.2
BART-large (Lewis et al., 2020a)	26.5	570	26.7	27.4
Retrieval-only models				
Dense Retriever (Lewis et al., 2021a)	26.7	-	28.9	-
DensePhrases (Lee et al., 2021)	40.9	18	50.7	-
RePAQ-base (Lewis et al., 2021b)	40.9	1400	39.7	29.4
RePAQ-large (Lewis et al., 2021b)	41.2	1100	-	-
RePAQ-xlarge (Lewis et al., 2021b)	41.5	800	41.3	-
Retrieval-augmented models				
REALM (Guu et al., 2020)	40.4	-	55.8	40.7
DPR (Karpukhin et al., 2020)	41.5	2.7	57.9	42.4
QAMAT (Chen et al., 2022)	44.7	240*	48.0	39.4
RePAQ rerank (Lewis et al., 2021b)	45.7	55	48.9	37.6
RAG (Lewis et al., 2020b)	44.5	9.6	56.8	45.2
FiD-base (Izacard and Grave, 2021)	48.2	3.7	65.0	32.4
FiD-large (Izacard and Grave, 2021)	51.4	1.4	67.6	-
Ours				
EMAT-FKSV	44.3	1000	44.4	36.7
EMAT-SKSV	43.3	1200	43.7	33.2

Still trails the best open book methods: QG has limited **recall**

Fast key slow value
Slow key slow value

High-recall question generation

The PAQ question generation process

1. Generate answer strings in each document *d*: *a ~ Pr(a | d)*
2. Generate questions conditioned on the answer and document: *q ~ Pr(q|a, d)*
3. Filter the questions to make sure they make sense as “open book” questions
   76% of questions are discarded as having multiple “interpretations”

d = “The Michigan Wolverines men’s basketball team have played their home games at Crisler Center since 1967. Crisler also became the home stadium of the women’s basketball team since 1974... The home stadium of the Wolverines football team is the Michigan Stadium.”

a = “Crisler Center”

a = “1967”

a = ...

q = “where does the men’s basketball team play their home games”

The PAQ question generation process

1. Generate answer strings in each document *d*: *a ~ Pr(a | d)*
2. Generate questions conditioned on the answer and document: *q ~ Pr(q|a, d)*
   - Use a question generation model trained on NQ
3. Filter the questions to make sure they make sense as “open book” questions

AmbigQA dataset (Min et al, 2021) is an annotated **subset** of NQ questions

- identified as ambiguous in an open-book context
- and edited by crowdworkers to be **unambiguous**

q = “where does the men’s basketball team play their home games”

q = “where does the **University of Michigan** men’s basketball team play their home games **since 1967**”

The SIXPAQ* question generation process

Haitian Sun, paper to appear soon!

1. Generate answer strings in each document *d*: *a ~ Pr(a | d)*
2. Generate questions conditioned on the answer and document: *q ~ Pr(q|a, d)*
    - Use a question generation model trained on ~~NQ~~ AmbigQA crowdworker-edited questions, with a single “interpretation”

~~3. Filter the questions to make sure they make sense as “open book” questions~~

Other differences:

- Generate from all passages
    - PAQ generates from a subset
- Generate one question from each answer
    - PAQ generates two, often redundant
- Filter questions generated from *a, d* if *a* is not the best answer to *q* in *d*
    - PAQ does this also

Result:

- 21m passages
- 137M questions
- 127M answers

vs PAQ’s 60M questions after filtering

*Single Interpretation eXtensions to Probably Asked Questions

SIXPAQ coverage

	AmbigQA		WebQSP		NQ	
	Ans	QA	Ans	QA	Ans	QA
PAQ	83.0	45.1	81.9	48.5	89.9	70.3
SIXPAQ	89.2	79.3	89.9	82.5	92.0	85.6

Table 1: Recall of answers of SIXPAQ on AmbigQA (dev), WebQuestionsSP (test) and NQ (dev). "Ans" represents the recall of answers by matching answer strings (after normalization steps). "QA" is evaluated on 100 randomly selected question by human annotators by checking whether questions of the matched answers in the database are actually relevant.

For 89.9% of NQ questions *q* some PAQ question *q'* has the same answer as *q*.

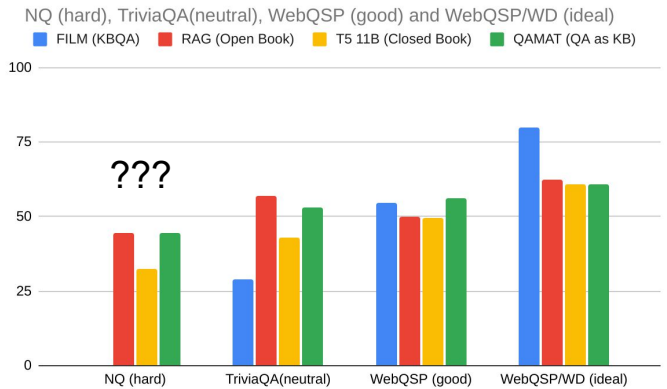
For 70.3% of NQ questions *q* some PAQ question *q'* has the same answer as *q* **and is semantically the same as *q***

*Single Interpretation eXtensions to Probably Asked Questions

SIXPAQ coverage

	AmbigQA		WebQSP		NQ	
	Ans	QA	Ans	QA	Ans	QA
PAQ	83.0	45.1	81.9	48.5	89.9	70.3
SIXPAQ	89.2	79.3	89.9	82.5	92.0	85.6

Table 1: Recall of answers of SIXPAQ on AmbigQA (dev), WebQuestionsSP (test) and NQ (dev). "Ans" represents the recall of answers by matching answer strings (after normalization steps). "QA" is evaluated on 100 randomly selected question by human annotators by checking whether questions of the matched answers in the database are actually relevant.

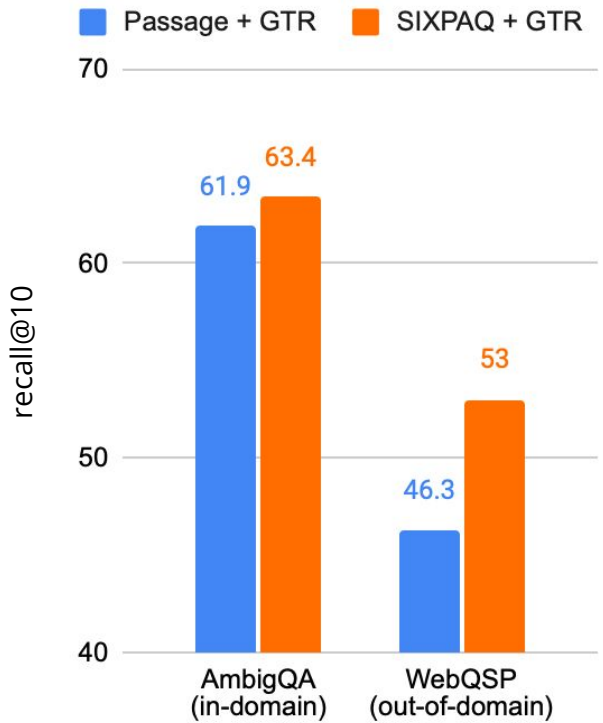


*Single Interpretation eXtensions to Probably Asked Questions

SIXPAQ applications

Question-based retrieval

Retrieve passages *indirectly* by matching asked questions *q* to generated questions *q'* (and then taking associated passage)



*Single Interpretation eXtensions to Probably Asked Questions

SIXPAQ applications

Long-form QA for ambiguous questions

Dataset: ASQA [Stelmakh et al. 2022]

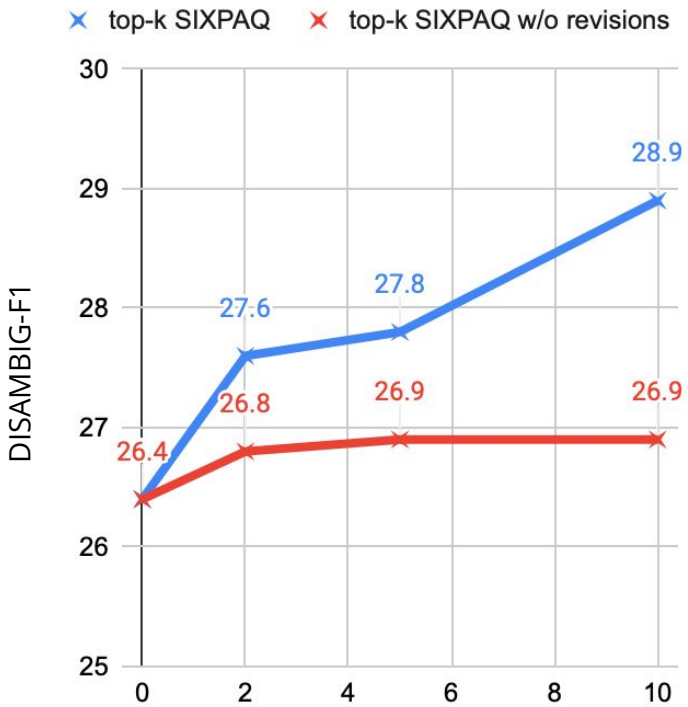
Q: who wrote the sound of silence

A: “The original artist of the song sound of silence released in 1966 is Paul Simon and Art Garfunkel. In 2016, Australian recording artist Dami Im recorded a different song by the same name. The album …”

Input: an AmbigQA question with

- Top-5 retrieved Wikipedia passage, plus
- Top-k most similar SIXPAQ questions

Metric: DISAMBIG-F1



*Single Interpretation eXtensions to Probably Asked Questions

Recap:
Questions, Answers, and Entities as a KB

Recap and conclusions

- 40 years of work has mostly solved KBQA
  - LLM-based semantic parsers and a KB inference engine work well
- …but broad-coverage KBs can be *misaligned* with information needs
  - Building KBs data-first means that some things users ask about will get missed
- Text-based open book QA methods still are challenged by
  - Multihop reasoning
  - Providing multiple answers
  - …
- Closed book QA methods with LLMs still lag in performance and in justifying answers

Recap and conclusions

- QA based KBs are a plausible bridge from structured to unstructured knowledge
  - can be easily linked to structured KBs (via entity linking on original text)
  - are made of modular, easily composed units (like a standard KB)
  - are useful as external memory for LLMs
    - for both short answers (NQ, TQA, …) and long-answer tasks
  - can be efficient as external memory for LLMs
  - have good coverage for user questions (latest: mid-80s for NQ)

Thanks for your attention!