CompMix: A Benchmark for Heterogeneous Question Answering

Philipp Christmann, Rishiraj Saha Roy, Gerhard Weikum

Max Planck Institute for Informatics
Saarbrücken, Germany

WWW 2024, Singapore
Heterogeneous question answering

**QA system**
- **Text**
  - Who was fouled before the first penalty in the 2022 FIFA final?
  - Ángel Di María

**QA system**
- **Tables**
  - Argentina’s ball possession in the 2022 WC final?
  - 59%
Heterogeneous question answering

Which team was behind by two goals but still won a FIFA final?

⇒ QA systems operating over heterogeneous sources as a result
⇒ Integrate multiple sources...
⇒ ...for broader answer coverage
Heterogeneous question answering

In which stadium was the 2022 soccer world cup final played?

⇒ QA systems operating over **heterogeneous sources as a result**
⇒ Integrate **multiple** sources...
⇒ ...for **broader answer coverage**
⇒ ...to leverage answer redundancy
Benchmarks for heterogeneous QA

Existing benchmarks
× Questions not human generated
× Span only two sources
× Domain-specific
× ...

✗

✗

✗

✗...
Benchmarks for heterogeneous QA

Existing benchmarks

- Questions not human generated
- Span only two sources
- Domain-specific
- ...

⇒ Methods for heterogeneous QA typically consider source-specific benchmarks, and artificially weaken one of the input sources (e.g. dropping 50% of the KB)
CompMix aims to fill this gap

Existing benchmarks
✗ Questions not human generated
✗ Span only two sources
✗ Domain-specific
✗ ...

CompMix benchmark (ours)
✓ Questions crowdsourced
✓ Span four kinds of sources (KB, text, tables, infoboxes)
✓ Covers five different domains (books, movies, music, tv series, soccer)
✓ ...

⇒ Collate completed questions from ConvMix (dataset for conversational QA)
⇒ CompMix has been used as a benchmark already
CompMix: Key statistics

★ 9,410 questions

★ 9.19 words per question
★ 2.17 words per answer
★ 5,413 entities

2,511 long-tail entities (<50 KB-facts)
CompMix: Complex phenomena

Which movie is longer, Hamlet or Gone with the Wind?

Which soccer player scored the most number of goals in the UEFA Euro 2004 tournament?

Author of the book To Kill a Mockingbird?

Who was the kit manufacturer of Chelsea Football Club from 1981 to 1983?
CompMix: Complex phenomena

- **Count**
  - How many matches has João Félix played for Portugal in 2019?
    - Tables

- **Ordinal**
  - Where did the Uruguay national football team play their first recorded match?
    - Text

- **Simple**
  - In what year was André Jardine born?
    - KB / Infoboxes / Text
Answer presence per source

How often the answer is present for any of the sources / source combinations

Answer presence

1.0
0.5
0.0

Kb  Text  Tables  Infoboxes  Kb+Text  Kb+Tables  Kb+Infoboxes  Text+Tables  Text+Infoboxes  Tables+Infoboxes  All four sources
Answer presence per source

Answer presence low for any single source

Single source
Answer presence per source

Answer presence low for any single source

Multiple sources
Answer presence per source

Incorporating heterogeneous sources required

All in!
Answer redundancy

In how many information sources each answer can be found

- Answer found in 1 source: 15.7%
- Answer found in 2 sources: 16.8%
- Answer found in 3 sources: 34.1%
- Answer found in all sources: 19.9%
Answer redundancy

For many questions, answer can be found in $\geq 3$ different sources.

$\Rightarrow$ Leveraging answer redundancy can pay off!
Experimental setup

- Experiments with diverse range of QA systems
  - **Generative LLMs**: GPT-3 (text-davinci-003)
  - **Heterogeneous QA**: UNIQORN, Unik-QA, EXPLAIGNN

- **Metrics**
  - Precision at 1 (**P@1**)
  - Mean reciprocal rank
  - Hits at 5
Evaluation with CompMix

Challenging dataset for existing heterogeneous QA methods, and LLMs

Precision at 1

0.0
0.5
1.0

Heterogeneous QA

LLM

CONVINSE [Christmann et al., SIGIR 2022]
UniK-QA [Oğuz et al., NAACL 2022]
EXPLAIGNN [Christmann et al., SIGIR 2023]
GPT-3 (text-davinci-003) [Brown et al., NeurIPS 2020]
Failure cases

None of the methods was able to answer these!

- Who played as adult Pi Patel in Life of Pi movie?
- Who were the twin brothers who played soccer for Manchester United?

Require judicious combination of heterogeneous sources:
- Textual (“adult Pi Patel” / ”twin brothers”)
- Structured (movie cast / club membership)
Desiderata in heterogeneous QA

Desiderata:
★ Integration of heterogeneous sources
★Grounding the answer to specific evidence
★ Traceability of the provided answer
Desiderata:
★ Integration of heterogeneous sources
★ Grounding the answer to specific evidence
★ Traceability of the provided answer

Demo at: https://explaignn.mpi-inf.mpg.de
CompMix – Walkthrough

★ Download from website

https://qa.mpi-inf.mpg.de/compmix

★ Load from HuggingFace datasets

```python
from datasets import load_dataset

dataset = load_dataset("pchristm/CompMix")
```
CompMix – Walkthrough

array [4966]
▼ 0 {8}
  question_id : 3642
  question : What is the genre of the tv series High Seas?
  domain : tvseries
▼ entities [1]
  ▼ 0 {2}
    id : Q59591953
    label : High Seas
▼ answers [1]
  ▼ 0 {2}
    id : Q186424
    label : detective fiction
    answer_src : kb
    answer_text : detective fiction
CompMix – Walkthrough

array [4966]
  ▼ 0 {8}
    question_id : 3642
    question : What is the genre of the tv series High Seas?
    domain : tvseries  Domain
  ▼ entities [1]
    ▼ 0 {2}
      id : Q59591953
      label : High Seas
  ▼ answers [1]
    ▼ 0 {2}
      id : Q186424
      label : detective fiction
      answer_src : kb
      answer_text : detective fiction
CompMix – Walkthrough

array [4966]

   ▼ 0 {8}
       question_id: 3642
       question: What is the genre of the tv series High Seas?
       domain: tvseries

       ▼ entities [1]
           ▼ 0 {2}
               id : Q59591953
               label: High Seas

       ▼ answers [1]
           ▼ 0 {2}
               id : Q186424
               label: detective fiction
               answer_src : kb
               answer_text : detective fiction
CompMix – Walkthrough

array [4966]

▼ 0 {8}
  question_id: 3642
  question: What is the genre of the tv series High Seas?
  domain: tvseries

▼ entities [1]
  ▼ 0 {2}
    id: Q59591953
    label: High Seas

▼ answers [1]
  ▼ 0 {2}
    id: Q186424
    label: detective fiction
    answer_src: kb
    answer_text: detective fiction

Answer:
- Text
- Wikidata ID
- Answer source
  (used by crowdworker)
Take-aways

★ New CompMix dataset for heterogeneous QA
  ★ 9,410 questions
  ★ Questions generated by humans

★ Integrating heterogeneous sources inherently required (KB, text, tables, infoboxes)
★ Existing methods merely answer 50% of questions correctly

Download CompMix at: https://qa.mpi-inf.mpg.de/compmix