A Benchmark for Temporal Question Answering with Implicit Time Constraints

Zhen Jia\textsuperscript{1}, Philipp Christmann\textsuperscript{2}, Gerhard Weikum\textsuperscript{2}

\textsuperscript{1}Southwest Jiaotong University, Chengdu, China
\textsuperscript{2}Max Planck Institute for Informatics, Saarbrücken, Germany

TempWeb @ WWW 2024, Singapore
Outline

★ Motivation
★ Construction of TIQ Benchmark
★ Characteristics of TIQ Benchmark
★ Experiments
★ Conclusion
Outline

★ Motivation
★ Construction of TIQ Benchmark
★ Characteristics of TIQ Benchmark
★ Experiments
★ Conclusion
Temporal questions

Which football club did Messi join in 2023?

⇒ Many user questions are time-sensitive
⇒ Temporal condition can be explicitly given ("in 2023")
Implicit questions

Which football club did Messi join after Paris Saint-Germain?

⇒ Natural to provide temporal condition implicitly
Implicit questions

Which football club did Messi join after Paris Saint-Germain?

- Natural to provide temporal condition implicitly
- QA system has to understand temporal condition to answer question
- Challenging for existing QA systems
Existing benchmarks for temporal QA

☆ Most benchmarks have **only few implicit questions**
  ☆ TempQuestions (209), TimeQuestions (1,476), TempQA-WD (154)

☆ Some have more implicit questions, but **limited** in question **intents**
  ☆ CronQuestions, TempReason: Both **template-based** and questions derived from **only 5/10 KB-relations**
Existing benchmarks for temporal QA

☆ Most benchmarks have only few implicit questions
  ☆ TempQuestions (209), TimeQuestions (1,476), TempQA-WD (154)

☆ Some have more implicit questions, but limited in question intents
  ☆ CronQuestions, TempReason: Both template-based and questions derived from only 5/10 KB-relations

<table>
<thead>
<tr>
<th>CronQuestions</th>
<th>TempReason</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Member of sports team</td>
<td>1. Member of sports team</td>
</tr>
<tr>
<td>2. Position held</td>
<td>2. Position held</td>
</tr>
<tr>
<td>3. Award received</td>
<td>3. Employer</td>
</tr>
<tr>
<td>4. Spouse</td>
<td>4. Political party</td>
</tr>
<tr>
<td>5. Employer</td>
<td>5. Head coach</td>
</tr>
<tr>
<td></td>
<td>6. Educated at</td>
</tr>
<tr>
<td></td>
<td>7. Chairperson</td>
</tr>
<tr>
<td></td>
<td>8. Head of government</td>
</tr>
<tr>
<td></td>
<td>9. Head of state</td>
</tr>
<tr>
<td></td>
<td>10. Owned by</td>
</tr>
</tbody>
</table>
Existing benchmarks for temporal QA

☆ Most benchmarks have **only few implicit questions**
  ☆ TempQuestions (209), TimeQuestions (1,476), TempQA-WD (154)

☆ Some have more implicit questions, but **limited** in question **intents**
  ☆ CronQuestions, TempReason: Both **template-based** and questions derived from **only 5/10 KB-relations**

☆ Questions are derived from **single source only**
  ☆ TempQuestions, TimeQuestions, TempQA-WD, TempTabQA, CronQuestion, TempReason
TIQ: Temporal Implicit Questions

★ New benchmark for temporal QA with 10,000 implicit questions

★ Questions derived from multiple knowledge sources
  ★ Wikipedia text
  ★ Wikipedia infoboxes
  ★ Wikidata Knowledge Base
TIQ: Temporal Implicit Questions

- Derived using automatic construction pipeline
- Configurable among various dimensions
  - Temporal scope of questions
  - Domain diversity of topic entities
  - Ratio of prominent vs. long-tail entities
  - Question complexity
  - Total number of questions
  - ...

★
Outline

★ Motivation
★ Construction of \textit{TIQ} Benchmark
★ Characteristics of \textit{TIQ} Benchmark
★ Experiments
★ Conclusion
Construction of OTIQ benchmark

**Key idea #1:** Implicit questions consist of **main question** and **implicit constraint**

⇒ Initialize parts individually, based on different information snippets

---

"Messi joined American club Inter Miami in July 2023."

"Lionel Messi, member of sports team, Paris Saint-Germain, start time, August 2021, end time, 30 June 2023"

---

**Main question**

**Implicit constraint part**
Key idea #2: Start from Wikipedia year pages

⇒ Provide salient information

2024

Wikipedia year pages

List of notable events (with dates)
Construction of OTIQ benchmark

Sample topic entities from Wikipedia year pages
Sample topic entities from Wikipedia year pages

Retrieve information snippets from Wikipedia (text and infoboxes), and Wikidata (KB)
Construction of OTIQ benchmark

Sample topic entities from Wikipedia year pages

Retrieve information snippets from Wikipedia (text and infoboxes), and Wikidata (KB)

Connect snippets with temporal relation
Construction of OTIQ benchmark

Sample topic entities from Wikipedia year pages

Retrieve information snippets from Wikipedia (text and infoboxes), and Wikidata (KB)

Connect snippets with temporal relation

Rephrase the pseudo-question into a natural question

```
What sports hall of fame Alan Page he was inducted into during Alan Page Associate Justice of the Minnesota Supreme Court In office?
```

```
“What hall of fame did Alan Page become a member of while serving as Associate Justice of the Minnesota Supreme Court?”
```
Outline

★ Motivation
★ Construction of TIQ Benchmark
★ Characteristics of TIQ Benchmark
★ Experiments
★ Conclusion
Alicia Keys followed up her debut with *The Diary of Alicia Keys*, which was released in December 2003.

Norah Jones, award received, Grammy Award for Best New Artist, follows, Alicia Keys, point in time, 2003

What album did Alicia Keys release when Norah Jones won the Grammy Award for Best New Artist?
Chris Brown, His fifth album, Fortune, released in 2012, also topped the Billboard 200.

“Which album released by Chris Brown topped the Billboard 200 when he was performing in Sydney?”
Example: #3

**Topic entity**: Clarence Andrew Cannon

**Evidence**: “Clarence Andrew Cannon, occupation, teacher, start time, 1904, end time, 1908”

“Clarence Cannon, He earned an LL.B. and joined the bar in 1908.”

**Pseudo-question**: What position Clarence Andrew Cannon occupation before Clarence Cannon

He earned an LL.B. and joined the bar

**Question**: “What was Clarence Andrew Cannon's occupation before becoming a lawyer?”
Example: #4

What television series was Hulk Hogan starring in when he signed with World Championship Wrestling (WCW)?
TIQ: Key statistics

★ 10,000 implicit questions
  6,000 train
  2,000 dev
  2,000 test

★ Derived from 10,000 topic entities

★ Temporal relations
  "before": 14%
  "during": 66%
  "after": 20%
**TIQ: Balance of entity popularity**

★ Balanced fraction of long-tail entities vs. prominent entities
  - long-tail entity: < 20 KB-facts
  - prominent entity: > 500 KB-facts

★ 2,542 long-tail topic entities

★ 2,613 prominent topic entities

⇒ *Actively controlled* when sampling topic entities
⇒ *Challenging test bed for LLMs*
TIQ: Diverse source combinations

⇒ Source combinations balanced as well
⇒ Not actively controlled
Further statistics (more in the paper)

★ Temporal values
  12,094 years
  538 months
  7,992 dates

★ Year pages: [1801, 2025]

★ 2.45 entities per question

★ 17.96 words per question
Outline

★ Motivation
★ Construction of ❌TIQ Benchmark
★ Characteristics of ❌TIQ Benchmark
★ Experiments
★ Conclusion
Experimental setup

★ Experiments with diverse range of QA systems
  ★ Generative LLMs: InstructGPT, GPT-4
  ★ Heterogeneous QA: UNIQORN, UniK-QA, EXPLAIGNN
  ★ Temporal QA: EXAQT, FAITH (WWW2024)
Which football club did Messi join after Paris Saint-Germain?
Which football club did Messi join after Paris Saint-Germain?

When did Lionel Messi play for Paris Saint-Germain?

Intermediate question

FAITH

August 2021 – July 2023

Intermediate answers

Question entity: Messi
Question relation: which football club did join
Expected answer type: football club
Temporal relation: after
Temporal category: implicit
Temporal value: August 2021 – July 2023

Time-aware Structured Frame

Temporal Question Understanding
Temporal Question Understanding

Which football club did Messi join after Paris Saint-Germain?

Question entity: Messi
Question relation: which football club did join
Expected answer type: football club
Temporal relation: after
Temporal category: implicit
Temporal value: August 2021 – July 2023

Capture temporal information need of user intent in a structured manner
Which football club did Messi join after Paris Saint-Germain?

Capture temporal information need of user intent in a structured manner

Temporal Question Understanding
- Implicit Question Resolver

Time-aware Structured Frame

Faithful Evidence Retrieval
- Temporal Pruning

Obtain set of evidences from heterogeneous sources satisfying the temporal conditions

Faithful Evidence

KB
Text
Tables
Infoboxes
Which football club did Messi join after Paris Saint-Germain?

Temporal Question Understanding
  Implicit Question Resolver
  Time-aware Structured Frame

Faithful Evidence Retrieval
  Temporal Pruning
  Faithful Evidence

Explainable Answering
  Answer + supporting evidence

Capture temporal information need of user intent in a structured manner

Obtain set of evidences from heterogeneous sources satisfying the temporal conditions

Derive the answer from the faithful evidences
Experimental setup

★ Experiments with diverse range of QA systems
  ★ **Generative LLMs:** InstructGPT, GPT-4
  ★ **Heterogeneous QA:** UNIQORN, UniK-QA, EXPLAIGNN
  ★ **Temporal QA:** EXAQT, FAITH (WWW2024)

★ Metrics
  ★ Precision at 1 (**P@1**)
  ★ Mean reciprocal rank
  ★ Hits at 5
Experimental results

Implicit questions are very challenging for LLMs

Precision at 1

<table>
<thead>
<tr>
<th></th>
<th>LLM</th>
<th>Heterogeneous QA</th>
<th>Temporal QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>InstructGPT</td>
<td>0.237</td>
<td>0.236</td>
<td>0.232</td>
</tr>
<tr>
<td>GPT-4</td>
<td>0.236</td>
<td>0.425</td>
<td>0.491</td>
</tr>
<tr>
<td>UNIQORN</td>
<td>0.446</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniK-QA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXPLAIGNN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXAQT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAITH</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experimental results

### Precision at 1

- **LLM**
  - InstructGPT: 0.237
  - GPT-4: 0.236
  - UNIQORN: 0.236
  - UniK-QA: 0.236
  - EXPLAIGNN: 0.425
  - EXAQT: 0.446
  - FAITH: 0.232

- **Heterogeneous QA**
  - InstructGPT: 0.425
  - GPT-4: 0.446
  - UNIQORN: 0.425
  - UniK-QA: 0.425
  - EXPLAIGNN: 0.491
  - EXAQT: 0.491
  - FAITH: 0.491

- **Temporal QA**
  - InstructGPT: 0.491
  - GPT-4: 0.491
  - UNIQORN: 0.491
  - UniK-QA: 0.491
  - EXPLAIGNN: 0.491
  - EXAQT: 0.491
  - FAITH: 0.491

*External grounding helps*
Experimental results

Best performance with FAITH, which explicitly resolves implicit questions

- LLM
- Heterogeneous QA
- Temporal QA

Precision at 1

- InstructGPT
- GPT-4
- UNIQORN
- UniK-QA
- EXPLAIGNNN
- EXAQT
- FAITH
Experimental results

Precision at 1

Performance below 50%: plenty of room for improvement!

- **LLM**: 0.237, 0.236
- **Heterogeneous QA**: 0.236, 0.425
- **Temporal QA**: 0.446, 0.491

- InstructGPT
- GPT-4
- UNIQORN
- UniK-QA
- EXPLAIGNN
- EXAQT
- FAITH
Common failure cases

★ 66.05%: no LLM able to answer

What club did Francisco Gento manage after Palencia?
Temporal constraint ignored

★ 34.80%: no heterogeneous QA method able to answer

Which university did Thomas Hunt Morgan attend after receiving his Bachelor of Science degree?
Temporal constraint misunderstood

★ 41.50%: no temporal QA method able to answer

What was the first line of the Saint Petersburg Metro to open when it began operations?
Evidence scoring issues

★ 16.65%: no method able to answer

During Marcone’s Senior career, what football club did he play for while representing the Qatar Olympic team?
Other complexities (long-tail entities, “during” and “while” in single question,...)
Outline

★ Motivation
★ Construction of TIQ Benchmark
★ Characteristics of TIQ Benchmark
★ Experiments
★ Conclusion
Meta-data

★ Question
★ Answer (text, Wikidata KB, Wikipedia)
★ Information snippets (text, source, temporal values)
★ Question derivation (topic entity, temporal relation, pseudo-question)
★ Question entities
★ Question reference date

question: What album did Alicia Keys release when Norah Jones won the Grammy Award for Best New Artist?
pseudo_question: What album Alicia Keys Keys followed up her debut with which was released, during, Norah Jones award received Grammy Award for Best New Artist follows Alicia Keys

- evidence [2]
  - temporal_relation: during
- topic_entity [5]
- question_entities [6]
- answer [1]
  - id: 17
  - question_creation_date: 2023-07-15
  - data_set: train
Conclusion

★ Automatic method to construct implicit questions
  ★ Highly configurable
  ★ Accesses multiple sources
  ★ Diverse information needs

★ **tiQ** benchmark: **Temporal Implicit Questions**
  ★ New benchmark with 10,000 implicit temporal questions
  ★ Derived from Wikipedia text, Wikipedia infoboxes, and Wikidata KB
  ★ Challenging testbed for temporal QA systems
  ★ Useful beyond QA?

★ Code and data: [qa.mpi-inf.mpg.de/tiq](http://qa.mpi-inf.mpg.de/tiq)

Thank you!