SECTION 9 · 10 min

Emerging Topics & Open Challenges

Where the field must go next to build temporally aware, trustworthy QA systems.

Presenter: Adam Jatowt

09

Dynamic Temporal Knowledge Management

Facts change constantly, yet TQA systems tend to rely on static corpora — while an update can silently break many related facts.

The Challenges

- Static corpora can't **keep up** with **fast-evolving information**.
- Updating one fact may disrupt dependent facts — the **temporal propagation problem**.
- Architectures lack the **modularity to edit and reason over such dependencies** in real time.
- Conflicting temporal evidence is hard to reconcile.

Example:

Q: “*What policies were enacted during the leader's presidency?“*

If the leader's term-end date changes, every dependent question ("before they left office") must update, too — current systems fail to propagate this.

Future directions

- Move from static corpora to **update-aware temporal knowledge**.
- Propagate **updates** to related facts, questions, and answers.
- Represent facts with **time, provenance, and dependencies**.
- Enforce consistency across **timelines** and **multi-turn interactions**.
- **Reconcile conflicting evidence** as knowledge evolves.

Temporal Uncertainty & Confidence

Time information is often uncertain, approximate, disputed, or incomplete, yet most QA systems assume exact temporal facts.

The Challenges

Many events do not have a single definitive timestamp.

- Historical events often have fuzzy boundaries or sources may disagree on dates or durations.
- Relative expressions ("early 1990s", "around 1200 BCE") are inherently imprecise.
- Small temporal errors may compound during multi-step reasoning.

Example:

Q: *"What happened between the fall of Rome and the Renaissance?"*

Both events have debated and region-dependent boundaries, making exact temporal reasoning difficult.

Future directions

- **Probabilistic representations** of uncertain dates & durations
- **Confidence propagation** through multi-step reasoning
- **Conflict-aware reasoning** over disagreeing sources
- **Uncertainty-aware generation** that communicates confidence
- **Evaluation that rewards appropriate uncertainty**, not false precision

Implicit Temporal Intent Understanding

Many questions hide their intended timeframe, and the same question may have different answers depending on when it's asked.

The Challenges

- Time requirements are **implied**, not stated.
- Intent depends on **conversational context, cultural assumptions, domain conventions**, and **user perspective**.
- Models **over-rely** on explicit temporal expressions.
- Systems miss the ambiguity or default to the most recent reading.

Example:

Q: "*What caused the economic crisis during Trump's presidency?*"

Could mean 2017–2021 events (e.g., COVID-19) or a more recent crisis — the temporal anchor is ambiguous.

Future directions

- **Contextual intent-detection** models
- **User**- and **domain**-aware temporal reasoning
- **Clarification strategies** for resolving ambiguity
- **Evaluation frameworks** for temporal intent detection

Temporally-Aware LLM Agents

LLM agents may reason well in general, but they sometimes lose the temporal thread of a conversation. They need memory and timeline mechanisms so that time references remain stable across turns.

The Challenges

- **Temporal hallucinations** — plausible but temporally wrong answers.
- Can't resolve context-dependent expressions ("last Tuesday," "since our last discussion").
- Transformers lack persistent temporal working memory.
- Turns treated independently → anchors drift over long interactions.

Example:

Q: “What did we decide last Tuesday?“

The agent can't anchor "last Tuesday" or carry temporal context forward, giving inconsistent answers.

Future directions

- Identify **temporal dependency chains** when facts change
- **Propagate updates** through related facts
- Maintain multi-hop temporal consistency
- **Reconcile conflicting evidence**; integrate temporal knowledge graphs

Diachronic & Synchronic Knowledge Integration

Many questions need both how things changed and what's true now — but systems treat these sources separately.

The Challenges

- Diachronic (change over time) and synchronic (snapshot) sources are handled in isolation.
- Requires alignment across temporal granularities and anchoring schemes.
- **Historical accounts and retrospective summaries can conflict**.
- Most models handle only one source type well.

Example:

Q: “How has unemployment changed since 2008, and what is the current rate?“

Needs long-term archive trends (diachronic) plus a recent statistic (synchronic) in one answer.

Future directions

- **Integrate** historical trajectories with current snapshots.
- **Align sources** across different temporal granularities.
- Combine archival, retrospective, and real-time evidence in **one answer.**
- **Detect and resolve conflicts** between past accounts and current knowledge.
- **Build unified representations** for evolving and present-state knowledge.

Multilingual & Multimodal Temporal QA

Temporal cues vary across languages and modalities, yet most systems assume English text.

The Challenges

- Non-Gregorian dates, varied formats, and cultural references.
- Temporal signals in images/video (seasonal imagery, handwritten timestamps).
- Limited cultural grounding and weak multimodal integration.

Example:

A lunar-calendar date in Arabic, or a video of a snowstorm implying winter — current systems fail to interpret these non-English / non-textual temporal signals.

Future directions

- **Multilingual** temporal taggers
- Temporally annotated **datasets** in **low-resource languages**
- **Cross-modal alignment** over text, images, and video
- **Culturally grounded temporal reasoning**

Evaluation & Benchmarking for Temporal Reasoning

Standard metrics miss temporal coherence — and benchmark/pretraining overlap may inflate scores.

The Challenges

- Accuracy / F1 / MRR / NDCG overlook **anchoring**, **ordering**, and **multi-hop consistency**.
- Evaluation often reduces to span accuracy or multiple choice.
- Benchmark–pre training contamination (Wikipedia/news in training data).
- Hard to separate genuine reasoning from memorization.

Example:

TimeQA and ArchivalQA have been derived from Wikipedia/news likely seen in pretraining, so high scores may reflect recall, not reasoning.

Future directions

- **Metrics** for temporal grounding and coherence
- Metrics sensitive to **ambiguous/conflicting time signals**
- Unified protocols for comparing temporal capabilities
- Temporal decontamination + out-of-distribution evaluation (RealTimeQA, FreshQA, Test of Time)

Domain-Specific Temporal Intelligence

Temporal reasoning is highly domain-dependent, yet most TQA systems are developed and evaluated on general-purpose corpora such as Wikipedia and news archives.

The Challenges

- Temporal semantics differ substantially across domains.
- **Domain-specific terminology and events require specialized knowledge.**
- Existing benchmarks fail to capture real-world temporal reasoning needs.

Future directions

- **Domain-specific temporal ontologies** and **event schemas**
- **Temporal foundation models adaptable across domains**
- **Specialized temporal retrievers** and RAG systems
- **Explainable temporal reasoning** for high-stakes decisions
- Benchmarks beyond news and Wikipedia

Example:

Healthcare:

Q: *How did the patient's condition evolve after starting immunotherapy?*

To answer correctly, the system must reason over:

- treatment timeline
- disease progression
- clinical events
- laboratory results

..not simply retrieve temporally relevant documents.

Legal

Q: Which regulation was in force before the 2018 amendment?

The answer depends on:

- effective dates
- amendment history
- legal dependencies
- jurisdiction-specific timelines