A Benchmark for Generalizable and Interpretable Temporal Question Answering over Knowledge Bases

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, Saswati Dana, Dinesh Garg, Achille Fokoue, G P Shrivatsa Bhargav, Dinesh Khandelwal, Srinivas Ravishankar, Sairam Gurajada, Maria Chang, Rosario Uceda-Sosa, Salim Roukos, Alexander Gray, Guilherme Lima, Ryan Riegel, Francois Luus, I Venkata Subramaniam

IBM Research

Abstract

Knowledge Base Question Answering (KBQA) tasks that involve complex reasoning are emerging as an important research direction. However, most existing KBQA datasets focus primarily on generic multi-hop reasoning over explicit facts, largely ignoring other reasoning types such as temporal, spatial, and taxonomic reasoning. In this paper, we present a benchmark dataset for temporal reasoning, TempQA-WD, to encourage research in extending the present approaches to target a more challenging set of complex reasoning tasks. Specifically, our benchmark is a temporal question answering dataset with the following advantages: (a) it is based on Wikidata, which is the most frequently curated, openly available knowledge base, (b) it includes intermediate SPARQL queries to facilitate the evaluation of semantic parsing based approaches for KBQA, and (c) it generalizes to multiple knowledge bases: Freebase and Wikidata. The TempQA-WD dataset is available at https://github.com/IBM/tempqa-wd.

1 Introduction

The goal of Knowledge Base Question Answering (KBQA) systems is to answer natural language questions by retrieving and reasoning over facts in Knowledge Base (KB). While reasoning in KBQA is evolving as an important research direction, most existing datasets and research in this area have primarily focused on one-triple questions (Bordes et al., 2015a) or multi-hop reasoning questions (Dubey et al., 2019; Berant et al., 2013) as illustrated in Table 1. Currently, there is a lack of approaches and datasets that address other types of complex reasoning, such as temporal and spatial reasoning. In this paper, we focus on a specific category of questions called temporal questions, where answering a question requires reasoning about points and intervals in time. For example, to answer the question What team did Joe Hart play for before Man City?, the system must retrieve or infer when Joe Hart played for Man City, when Joe Hart played for other teams, and which of the latter intervals occurred before the former.

| Category | Example |
|----------|---------|
| Single-Hop | Who directed Titanic Movie? |
| SPARQL: | select distinct ?a where { wd:Q45478 wdt:P57 ?a } |
| Multi-hop | Which movie is directed by James Cameron starring Leonardo DiCaprio? |
| SPARQL: | select distinct ?a where { ?a wdt:P57 wd:Q42574. ?a wdt:P161 wd:Q38111. } |
| Temporal | What team did Joe Hart play for before Man City? |
| SPARQL: | select distinct ?u where { wd:Q187184 p:P54 ?e1. ?e1 ps:P54 wd:Q50602. wd:Q187184 p:P54 ?e2. ?e2 ps:P54 ?u. ?e1 pq:P580 ?et1. ?e2 pq:P582 ?et2. filter (?et2 <= ?et1) order by desc (?et2) limit 1 |

Table 1: Examples of Single-hop, Multi-hop and Temporal reasoning questions on Wikidata.

Progress on Temporal KBQA is hindered by a lack of datasets that can truely assess temporal reasoning capability of existing KBQA systems. To the best of our knowledge, TempQuestions (Jia et al., 2018a), TimeQuestions (Jia et al., 2021), and CronQuestions (Saxena et al., 2021) are the only available datasets for evaluating purely this aspect. These have, however, a number of drawbacks: (a) These contain only question-answer pairs and not their intermediate SPARQL queries which could be useful in evaluating interpretability aspect of KBQA approaches based on semantic parsing (Yih et al., 2014); (b) unlike regular KBQA datasets (Dubey et al., 2019; Diefenbach et al., 2017b; Azmy et al., 2018) that can attest KBQA generality over multiple knowledge bases such as DBpedia, and Wikidata, these are suited for a single KB; (c) TempQuestions uses Freebase (Freebase) as the knowledge base, which is no longer maintained...
and was officially discontinued in 2014 (Freebase).

Our aim in this paper is to fill the above-mentioned gaps by adapting the TempQuestions dataset to Wikidata and by enhancing it with additional SPARQL query annotations. Having SPARQL queries for temporal dataset is crucial to refresh ground truth answers as the KB evolves. We choose Wikidata for this dataset because it is well structured, fast evolving, and the most up-to-date KB, making it a suitable candidate for temporal KBQA. Our resulting dataset thus contains parallel annotations (on both Wikidata and Freebase). This will help drive research towards development of generalizable approaches, i.e., those that could be easily be adaptable to multiple KBs. In order to encourage development of interpretable, semantic parsing-based approaches, we (1) annotate the questions with SPARQL queries and answers over Wikidata, and (2) annotate a subset of this dataset with intermediate representations for the entities, relations, λ-expression along with SPARQL needed to answer the question.

The main contributions of this work are as follows:

- A benchmark dataset, called TempQA–WD, for building temporal-aware KBQA systems on Wikidata with parallel annotations on Freebase, thus encouraging the development of approaches for temporal KBQA that generalize across KBs.
- SPARQL queries for all questions in our benchmark, with a subset of the data also annotated with expected outputs from intermediate stages of the question answering process, i.e., entity and relation linking gold outputs. The goal here is to encourage interpretable approaches that can generate intermediate outputs matching those fine-grained annotations.

2 Related Work

Over the years, many question answering datasets have been developed for KBQA, such as Free917 (Cai and Yates, 2013), SimpleQuestions (Bordes et al., 2015b), WebQuestions (Berant et al., 2013), QALD-9 (Usbeck et al., 2017), LC-QuAD 1.0 (Trivedi et al., 2017), and LC-QuAD 2.0 (Dubey et al., 2019). In Table 2, we compare each of these datasets across the following features: (a) underlying KB, including subsequent extensions, e.g. Wikidata (Diefenbach et al., 2017b) and DBpedia (Azmy et al., 2018) based versions of SimpleQuestions, as well as the Wikidata subset of WebQSP (Sorokin and Gurevych, 2018); (b) reasoning types that are emphasized in the dataset; (c) availability of SPARQL queries, entities, and relationships for intermediate evaluation; and (d) the use of question templates, which can often generate noisy, unnatural questions. As Table 2 shows, our dataset is distinguished from prior work in its emphasis on temporal reasoning, its application to both Freebase and Wikidata, and its annotation of intermediate representations and SPARQL queries.

The most relevant KBQA dataset to our work is TempQuestions (Jia et al., 2018a), upon which we base TempQA–WD, as described in Section 3. CronQuestions (Saxena et al., 2021) is another dataset where emphasis is on temporal reasoning. However, this dataset also provides a custom KB derived from Wikidata which acts as a source of truth for answering the questions provided as part of the dataset.

3 Dataset

TempQuestions (Jia et al., 2018a) was the first KBQA dataset intended to focus specifically on temporal reasoning. It consists of temporal questions from three different KBQA datasets with answers from Freebase: Free917 (Cai and Yates, 2013), WebQuestions (Berant et al., 2013) and ComplexQuestions (Bao et al., 2016). We adapt TempQuestions to Wikidata to create a temporal QA dataset that has three desirable properties. First, in identifying answers in Wikidata, we create a generalizable benchmark that has parallel annotations on two KBs. Second, we take advantage of Wikidata’s evolving, up-to-date knowledge. Lastly, we enhance TempQuestions with SPARQL, entity, and relation annotations so that we may evaluate intermediate outputs of KBQA systems.

Two previous attempts at transferring Freebase-QA questions to Wikidata are WebQSP-WD (Sorokin and Gurevych, 2018) and SimpleQuestions-WD(SWQ-WD) (Diefenbach et al., 2017a). SWQ-WD is single triple questions and in WebQSP-WD only answers are directly mapped to corresponding entities in Wikidata. However, as stated (Sorokin and Gurevych, 2018), one challenge is that not all Freebase answers can be directly mapped to entities in Wikidata. For example, the Freebase answer annotation for the question "When did Moscow burn?" is “1812 Fire
of Mosco”, despite the year being entangled with the event itself. In contrast, Wikidata explicitly represents the year of this event, with an entity for “Fire in Moscow” and an associated year of “1812”. Thus, a direct mapping between the two answers is not possible, as it would amount to a false equivalence between “1812 Fire of Mosco” and “1812”.

In order to address such issues, we enlisted a team of annotators to manually create and verify SPARQL queries, ensuring not only that the SPARQL formulation was correct, but that the answers accurately reflected the required answer type (as in the “Fire in Moscow” example above) and the evolving knowledge in Wikidata (as the Freebase answers from the original dataset may be outdated). Having SPARQL queries also facilitates intermediate evaluation of the approaches that use semantic parsing to directly generate the query or the query graph, increasing interpretability and performance in some cases (Sorokin and Gurevych, 2018).

Next, we give a brief overview of how Wikidata is organized, as some of its representational conventions impact our dataset creation process.

### 3.1 Wikidata

Wikidata\(^1\) is a free knowledge base that is actively updated and maintained. It has 93.65 million data items as of May, 2021 and continuously growing each day with permission for volunteers to edit and add new data items. We chose Wikidata as our knowledge base as it has many temporal facts with appropriate knowledge representation encoded. It supports reification of statements (triples) to add additional metadata with qualifiers such as start date, end date, point in time, location etc. Figure 1 shows an example of reified temporal information associated to entities Joe Hart and Manchester City. With such representation and the availability of up-to-date information, Wikidata makes it a good choice to build benchmark datasets to test different kinds of reasoning including temporal reasoning.

### 3.2 Dataset Details

Table 3 gives details of our new benchmark dataset. We took all the questions from TempQuestions dataset (of size 1271) and chose a subset for which we could find Wikidata answers. This subset has 839 questions that constitute our new dataset, TempQA-WD. We annotated this set with their corresponding Wikidata SPARQL queries and the derived answers. We also retained the Freebase answers from the original TempQuestions dataset effectively creating parallel answers from two distinct KBs. Additionally, we added question com-

---

\(^{1}\)https://www.wikidata.org/
Table 3: Benchmark dataset details. TempQuestions denote original dataset with Freebase answers (Jia et al., 2018a). The remaining two, i.e., testset and devset, are part of the benchmark dataset we created with Wikidata answers. In column titled Add’l Details: Set-A denote additional annotations that came along with the original TempQuestions dataset, {temporal signal, question type, data source}. Set-B denote {Wikidata SPARQL query, question complexity category}. Set-C denote ground truths of {AMR, \(\lambda\)-expression, Wikidata entities, Wikidata relation, Wikidata-specific \(\lambda\)-expression}.

3.2.1 Question Complexity Categorization

| Category (dev/Test) | Example | Example |
|---------------------|---------|---------|
| Simple (92/471)     | When was Titanic movie released? | SELECT ?a WHERE { wd:Q44578 wdt:P54 wd:Q50602. wd:Q187184 p:q54 ?e1. ?e1 ps:q54 wd:Q50602. ?e1 p:q54 ?s1. wd:Q187184 wdt:P54 ?e2. ?e2 ps:q54 ?s2. FILTER (?s2 <= ?s1) | order by desc (?s2) LIMIT 1 |
| Medium (71/154)     | who was the US president during cold war? | SELECT DISTINCT ?a WHERE { ?a wdt:P39 wd:Q11696. ?a p:P39 ?e1. ?e1 ps:Q39 wd:Q11696. ?e1 p:Q580 ?s1. ?e1 p:Q580 ?s1. wd:Q8683 wdt:P580 ?s2. wd:Q8683 wdt:P582 ?s2. FILTER (?s1 <= ?s2 && ?s2 <= ?s1) } |
| Complex (12/59)     | who was president of the us when douglas bravo was a teenager? | SELECT DISTINCT ?a WHERE { ?a p:Q39 ?e. ?e ps:Q39 wd:Q11696. ?e p:Q580 ?s1. ?e p:Q582 ?s2. wd:Q4095606 wdt:P569 ?x. bind ("((x + "P13Y")\(\times\) xsd:duration) as ?s2) bind ("((x + "P19Y")\(\times\) xsd:duration) as ?et2) FILTER (?s1 <= ?s2 && ?s2 <= ?s1) ) |
For evaluations in Section 4, we divided dataset TempQA-WD of size 839 into two parts, TempQA-WD-test of size 664 and TempQA-WD-dev of size 175. In fact, TempQA-WD-dev is that part of the dataset with fine-grained annotations. In this dataset, we also labeled questions with complexity category based on the complexity of the question in terms of temporal reasoning required to answer. Table 4 shows the examples for each category of complexity defined below.

1) Simple: Questions that involve one temporal event and need no temporal reasoning to derive the answer. For example, questions involving simple retrieval of a temporal fact or simple retrieval of other answer types using a temporal fact.

2) Medium: Questions that involve two temporal events and need temporal reasoning (such as overlap/before/after) using time intervals of those events. We also include those questions that involve single temporal event but need additional non-temporal reasoning.

3) Complex: Questions that involve two or more temporal events, need one temporal reasoning and also need an additional temporal or non-temporal reasoning like teenager or spatial or class hierarchy.

3.3 Lambda Calculus

In this section, we give a brief description of λ-calculus, since we use λ-expressions to logically represent the semantics of questions. λ-calculus, by definition, is considered the smallest universal programming language that expresses any computable function. In particular, we have adopted Typed λ-Calculus presented in (Zettlemoyer and Collins, 2012). In addition to the constants and logical connectives, we introduced some new temporal functions and instance variables to avoid function nesting. For example, consider the following question and its corresponding logical form:

**Question:** when was Barack Obama born?

**Logical Form:***\( \lambda t. \text{born}(b, \text{"Barack Obama"}) \land \text{interval}(t, b) \)*

Here, *b* is instance variable for event *born(b, "Barack Obama")* and *interval(t, b)* finds time for the event denoted by *b*. Variable *t* is unknown which is marked as λ variable.

**Temporal Functions:** We introduce interval, overlap, before, after, teenager, year; where interval gets time interval associated with event and overlap, before, after are used to compare temporal events. teenager gets teenager age interval for a person, and year return year of a date.

4 Evaluation

4.1 Evaluation Setup

We use SYGMA (Neelam et al., 2021) to generate the baseline results for TempQA-WD dataset. The baseline system is tuned with dev set of TempQA-WD, and evaluated on the test dataset.

4.2 Metrics

We use GERBIL (Usbeck et al., 2019) to compute performance metrics from the pairs of gold answers and system generated answers from the pipeline. We use standard performance metrics typically used for KBQA systems, namely macro precision, macro recall and F1.

4.3 Results & Discussion

Table 5 shows performance of SYGMA along with the performance of TEQUILLA (Jia et al., 2018b) (that uses Freebase as its underlying KB) on TempQA-WD test set. The accuracy numbers show that there is room for improvement and good scope for research on temporal QA (and complex QA in general) on Wikidata. Table 6 gives detailed report of performance for different categories of question complexity. Performance of TEQUILLA reflects how well Accu and Quint (the QA systems it used) are adapted to Freebase. This in addition to the fact that majority of the questions fall under simple questions category (471 simple vs 39 complex as shown in Table 4), it is able to achieve better accuracy on Freebase. Given relatively less focus of research on temporal QA on Wikidata so far, we believe our benchmark dataset would help accelerate more research and datasets in the future.

To gain more insights on the performance, we also did an ablation study of SYGMA using TempQA-WD dev set where impact of individual module on overall performance is evaluated. Table 7 shows the results. For example GT-AMR refers to the case where ground truth AMR is fed directly into λ-module. The table shows large jump in accuracy (in both the datasets) when fed with ground truth entities (GT-EL) and ground truth relations (GT-RL). This points to the need for improved entity linking and relation linking on Wikidata.
| System | Dataset   | Precision | Recall | F1  |
|--------|-----------|-----------|--------|-----|
| SYGMA  | TempQA-WD | 0.32      | 0.34   | 0.32|
| TEQUILA| TempQA-FB | 0.54      | 0.48   | 0.48|

Table 5: Performance of SYGMA

| System | Category | Precision | Recall | F1  |
|--------|----------|-----------|--------|-----|
| SYGMA  | Simple   | 0.39      | 0.37   | 0.38|
|        | Medium   | 0.16      | 0.13   | 0.13|
|        | Complex  | 0.38      | 0.38   | 0.38|
| TEQUILA| Simple   | 0.59      | 0.53   | 0.53|
|        | Medium   | 0.46      | 0.39   | 0.40|
|        | Complex  | 0.20      | 0.20   | 0.18|

Table 6: Category wise Performance

|       | TempQA-WD | Precision | Recall | F1  |
|-------|-----------|-----------|--------|-----|
| NO GT |           | 0.47      | 0.50   | 0.47|
| GT-AMR|           | 0.50      | 0.51   | 0.50|
| GT-λ  |           | 0.52      | 0.53   | 0.52|
| GT-EL |           | 0.60      | 0.62   | 0.60|
| GT-RL |           | 0.92      | 0.93   | 0.92|
| GT-KB-λ|          | 0.93      | 0.93   | 0.93|
| GT-SPARQL|         | 1.0       | 1.0    | 1.0 |

Table 7: Ablation Study on TempQA-WD

5 Conclusion

In this paper, we introduced a new benchmark dataset TempQA-WD for temporal KBQA on Wiki-data. Adapted from existing TempQuestions dataset, this dataset has parallel answer annotations on two KBs. A subset of this dataset is also annotated with output expected at intermediate stages of modular pipeline. Future extensions include improvement of the baseline approach for generalizable temporal QA.

References

Michael Azmy, Peng Shi, Jimmy Lin, and Ihab Ilyas. 2018. Farewell Freebase: Migrating the Simple-Questions dataset to DBpedia. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2093–2103, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2503–2514, Osaka, Japan. The COLING 2016 Organizing Committee.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015a. Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015b. Large-scale simple question answering with memory networks. CoRR, abs/1506.02075.

Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 423–433, Sofia, Bulgaria. Association for Computational Linguistics.

Dennis Diefenbach, Kamal Singh, and Pierre Maret. 2017a. Wdaqua-core0: A question answering component for the research community. In Semantic Web Evaluation Challenge, pages 84–89. Springer.

Dennis Diefenbach, Thomas Pellissier Tanon, Kamal Deep Singh, and Pierre Maret. 2017b. Question answering benchmarks for wikidata. In Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia, pages 69–78.

Freebase. Freebase (database). Information about Freebase database at https://en.wikipedia.org/wiki/Freebase_(database).

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jan- nik Strötgen, and Gerhard Weikum. 2018a. Tempquessions: A benchmark for temporal question answering. In Companion Proceedings of the Web Conference 2018, WWW ’18, page 1057–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jan- nik Strötgen, and Gerhard Weikum. 2018b. Tequila: Temporal question answering over knowledge bases. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM ’18, page 1807–1810, New York, NY, USA. Association for Computing Machinery.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In Proceedings of the 30th ACM International Conference on
Question answering over knowledge bases by leveraging semantic parsing and neuro-symbolic reasoning.

Sumit Neelam, Udit Sharma, Hima Karanam, Shahjith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, et al. 2021. Sygma: System for generalizable modular question answering over knowledge bases. *arXiv preprint arXiv:2109.13430*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. *arXiv preprint arXiv:2106.01515*.

Daniil Sorokin and Iryna Gurevych. 2018. Modeling semantics with gated graph neural networks for knowledge base question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3306–3317. Association for Computational Linguistics.

A. Talmor and J. Berant. 2018. The web as a knowledge-base for answering complex questions. In *North American Association for Computational Linguistics (NAACL)*.

Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. Lc-quad: A corpus for complex question answering over knowledge graphs. In *Proceedings of the 16th International Semantic Web Conference (ISWC)*, pages 210–218. Springer.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 2017. 7th open challenge on question answering over linked data (QALD-7). In *Semantic Web Evaluation Challenge*, pages 59–69. Springer International Publishing.

Ricardo Usbeck, Michael Röder, Michael Hoffmann, Felix Conrads, Jonathan Huthmann, Axel-Cyrille Ngonga Ngomo, Christian Demmler, and Christina Unger. 2019. Benchmarking question answering systems. *Semantic Web*, 10(2):293–304.

Wen-tau Yih, Xiaodong He, and Christopher Meek. 2014. Semantic parsing for single-relation question answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 643–648.

Luke S. Zettlemoyer and Michael Collins. 2012. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars.