Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2998–3007
Marseille, 20-25 June 2022

Knowledge Graph Question Answering Leaderboard: A Community Resource to Prevent a Replication Crisis

A. Perevalov*§¶, Xi Yan†¶, L. Kovriguina‡, L. Jiang†, A. Both§ and R. Usbeck†
*Anhalt University of Applied Sciences, †University of Hamburg,
‡Fraunhofer IAIS, §Leipzig University of Applied Sciences
aleksandr.perevalov@hs-anhalt.de, liubov.kovriguina@iais.fraunhofer.de,
andreas.both@htwk-leipzig.de, {xi.yan, longquan.jiang, ricardo.usbeck}@uni-hamburg.de
¶These authors contributed equally to this work.

Abstract

Data-driven systems need to be evaluated to establish trust in the scientific approach and its applicability. In particular, this is true for Knowledge Graph (KG) Question Answering (QA), where complex data structures are made accessible via natural-language interfaces. Evaluating the capabilities of these systems has been a driver for the community for more than ten years while establishing different KGQA benchmark datasets. However, comparing different approaches is cumbersome. The lack of existing and curated leaderboards leads to a missing global view over the research field and could inject mistrust into the results. In particular, the latest and most-used datasets in the KGQA community, LC-QuAD and QALD, miss providing central and up-to-date points of trust. In this paper, we survey and analyze a wide range of evaluation results with significant coverage of 100 publications and 98 systems from the last decade. We provide a new central and open leaderboard for any KGQA benchmark dataset as a focal point for the community - https://kgqa.github.io/leaderboard/. Our analysis highlights existing problems during the evaluation of KGQA systems. Thus, we will point to possible improvements for future evaluations.

Keywords: Evaluation Methodology, Knowledge Graph, Question Answering, Leaderboard, Replication Crisis

1. Introduction

Question Answering (QA) is a rapidly growing field in research and industry[1]. QA systems already deliver their potential into many real-world problems, e.g., (Mutabazi et al., 2021; Both et al., 2021; Diefenbach et al., 2021). These systems can be divided into two main paradigms (Jurafsky and Martin, 2018): IR-based that works over unstructured data, closely related to Machine Reading Comprehension and Retriever-Reader architecture) and Knowledge-Based (KBQA) which works over structured data, such as relational tables, specific data APIs, knowledge graphs (KGs). In this regard, Question Answering over Knowledge Graphs (KGQA) is of particular interest to this work.

Many different benchmarking datasets are used for evaluating KGQA systems. These datasets differ in the underlying knowledge graph (e.g., DBpedia (Auer et al., 2007) or Wikidata (Erxleben et al., 2014)), size order of magnitude (Fu et al., 2020), questions complexity (Saleem et al., 2017), multilingual support (Chandra et al., 2021), and many more dimensions. In the KGQA research community, several datasets have become a de facto standard for evaluation of such systems, such as the QALD (Usbeck et al., 2018) and LC-QuAD (Dubey et al., 2019) benchmark dataset series. As more and more researchers introduce new evaluation results using these well-known datasets, it becomes more challenging to follow the up-to-date state-of-the-art in the KGQA field. The related research fields such as IR-based QA and Knowledge Graph research community already have their own well-established and maintained leaderboards of the best solutions (SQuAD[2] (Rajpurkar et al., 2016), OGB[3] (Hu et al., 2020)). However, it is not the case for KGQA. This lack - in particular of curated leaderboards - leads to a missing global view over the research field. In turn, this could inject mistrust into result tables within publications when they are incomplete or lack a comparison to certain systems, as often required by reviewers. In particular, the latest and most-used datasets in the Semantic Web community, LC-QuAD and QALD, miss providing central points of trust such as leaderboards. In this paper, we analyze the publications of KGQA evaluations of the last decade. We evaluated 100 papers and 98 systems on 4 datasets focusing on the LC-QuAD and QALD series. Our results show that evaluation numbers are often consistent. Existing errors stem from minor differences in the data (e.g., gAnswer (Hu et al., 2018) on QALD-9 (Usbeck et al., 2018)) that seems to be rounding errors or inconclusive behavior. Finally, we discuss the consequences of our findings and will point to possible improvements for future evaluations.

Our contributions are as follows:

- We present the first, extensive evaluation analysis of the state of the research in KGQA.

[1] https://www.gartner.com/smarterwithgartner/2-megatrends-dominate-the-gartner-hype-cycle-for-artificial-intelligence-2020 (September 28, 2020)

[2] cf., The Stanford Question Answering Dataset leaderboard at https://rajpurkar.github.io/SQuAD-explorer/

[3] Open Graph Benchmark – is a collection of the benchmark datasets for machine learning over graphs: https://ogb.stanford.edu

- We provide a new central and open leaderboard for any KGQA benchmark dataset as a focal point for the community - https://github.com/KGQA/leaderboard. With pull requests, the community can easily enhance the leaderboard.
- We provide an up-to-date overview of all available demos or Web services for KGQA at the point of publication.

These contributions should help the scientific community to foster replication and cross-evaluation in the future.

In the following, we analyze related studies and approaches in Section 2. Afterward, we introduce the analyzed datasets and systems in Section 3. In Section 4, we describe our extensive state-of-the-art data and delve into its analysis. Next, we discuss possible interpretations and paths forward and end with a summary and outlook in Section 6.

2. Related Work

There are multiple approaches to tracking the progress of any research field. In machine learning and NLP, these approaches can be subdivided into *benchmarking frameworks* and *manual or semi-automatic reporting platforms*.

Today, benchmarking frameworks need to limit their scope to a subset of tasks to cover the necessary metrics and experiment types out-of-the-box. A general benchmarking framework, which works without writing code, does not exist. For KGQA, different *benchmarking* frameworks have been proposed. For example, GERBIL QA (Usbeck et al., 2019), can benchmark KGQA systems via their Web APIs in a FAIR way (Wilkinson et al., 2016). It also has an integrated leaderboard[4] which displays a summary of all experiments run via the platform. At the same time, this is the biggest downside – only experiments run via the platform are integrated. Thus, a realistic view depends on the adoption of the platform. This adoption seems to lack due to missing developer resources, which continuously update available systems and datasets. A different direction is followed by systems like https://github.com/AKSW/irbench or QALD-Gen (Singh et al., 2019), which provide command-line tools for benchmarking any KGQA system. However, the offline nature of these tools leads to offline results, i.e., the results might be used in papers but do not contribute to a trustworthy overview of the field of research.

Recently, *reporting platforms* gained popularity. They allow quick access to results, but either they are curated manually via a community or semi-automatically updated. https://nlpprogress.com/ is a famous community website launched by Sebastian Ruder. Regarding KGQA, the website's most recent information is 3 years old, possibly displaying the disinterest of the NLP community in semantic tasks. https://paperswithcode.com/ is another reporting platform run by Facebook AI research allowing to openly edit papers, code, datasets, methods, and evaluation tables. While this is of tremendous help for reproducibility, its results for KGQA are sparse. There is only one result for LC-QuAD 2 and QALD-9, and both are for relation extraction rather than Question Answering.

A promising semi-automatic approach is the Open Research Knowledge Graph (ORKG) (Auer et al., 2020). It allows the community to persistently annotate papers via smart tools with meta and evaluation data, e.g., https://www.orkg.org/orkg/paper/R6386/R6393 for QALD-6 data. However, the current adoption in the community does not go beyond prototypes provided by the ORKG team. A change might come with the European Open Science Cloud (EOSC) and the Nationale Forschungsdateninfrastruktur für Data Science und Künstliche Intelligence (German National Data Infrastructure for Data Science and AI). Those publicly funded initiatives strive to foster ecosystems like ORKG in the long term.

Finally, surveys can be viewed as reporting platforms. Different surveys have been published in the past decade focusing on a variety of topics such as challenges in general KGQA (Höffner et al., 2017), challenges in complex KGQA (Fu et al., 2020), core techniques of KGQA (Diefenbach et al., 2018) or neural network-based KGQA systems (Chakraborty et al., 2021). However, these are automatically outdated when published or focus only on a narrow subtopic.

Thus, there is the need for a central, dense, and open reporting platform focusing on KGQA, which provides trustworthy insights.

3. Benchmark Datasets and Systems

We surveyed 14 DBpedia-based KGQA benchmark datasets that were published in the last decade (cf., Section 3.1). In this paper, we consider 4 KGQA datasets for an in-depth analysis. Requirements for selecting a dataset include usage for the evaluation of different systems, availability in English, relying on DBpedia (primarily) or Wikidata (knowledge bases, which are still maintained), and cited above 5 times. Our goal was to make sure that the chosen QA datasets are: up-to-date, close to a real-world setting, can be manually evaluated, and are vastly studied. Note, we use benchmark datasets and datasets synonymous.

We took 98 QA systems into the consideration. They are collected manually from articles that include evaluation results on the considered benchmark datasets. The article search was conducted in two ways. First, we retrieved articles using a keyword search on Google Scholar[5]. Specifically, the selection criteria were: published after 2019, and titles satisfy: `['question`

[4] http://gerbil-qa.aksw.org/gerbil/overview

[5] http://scholar.google.com

`answering' AND ('semantic web' OR 'data web' OR 'web of data')]`. The second method is to extract all articles which cite the benchmark dataset from Google Scholar either as direct citation or as URL to the location of the dataset. We removed duplicates and manually extracted the QA systems evaluated or referred to in the articles. This resulted in 100 analyzed papers. Note, some systems are evaluated on a subset of the dataset or a dataset where the benchmark dataset is just a subset. We indicated such a difference in the leaderboard accordingly.

3.1. KGQA Datasets

The first dataset is QALD which is multilingual dataset challenge series. In QALD-8, there were 219 training question-answer pairs and 42 test data points respectively. It was the first edition to use GERBIL QA as a benchmarking platform (Usbeck et al., 2019). The newest instance – QALD-9 (Usbeck et al., 2018) – contains 558 questions incorporating information from the DBpedia knowledge base[6] where for each question the following is given: a textual representation in multiple languages, the corresponding SPARQL query (over DBpedia), the answer entity URI, and the answer type. The QALD series has a growing number of questions per edition and thus grows continuously in its expressiveness. The dataset has become a staple for many research studies in QA (e.g., (Höffner et al., 2017; Diefenbach et al., 2018)).

The second and third dataset is LC-QuAD. LC-QuAD (version 1) (Trivedi et al., 2017) is a Question Answering dataset with 5000 pairs of questions and its corresponding SPARQL query. LC-QuAD v2 (Dubey et al., 2019) is the follow-up dataset with 30.000 question-answer pairs to better suit novel machine learning approaches. The SPARQL queries are intended to be executed on DBpedia. LC-QuAD is widely used in the process of QA systems development (Singh et al., 2018; Dubey et al., 2018).

Other KGQA datasets are Free917 (Cai and Yates, 2013), WebQuestions (Berant et al., 2013), ComplexQuestions (Bao et al., 2016), SimplesQuestions (Bordes et al., 2015), GraphQuestions (Su et al., 2016), WebQuestionsSP (Yih et al., 2016), 30MFactoidQA (Serban et al., 2016), ComplexWebQuestions (Talmor and Berant, 2018), PathQuestion (Zhou et al., 2018), MetaQA (Zhang et al., 2018), TempQuestions (Jia et al., 2018), TimeQuestions (Jia et al., 2021), CronQuestions (Saxena et al., 2021), FreebaseQA (Jiang et al., 2019), Compositional Freebase Questions (CFQ) (Keysers et al., 2019), Compositional Wikidata Questions (CWQ) (Cui et al., 2021), RuBQ (Korablinov and Braslavski, 2020; Rybin et al., 2021), QALD-9-plus (Perevalov et al., 2022), GrailQA (Gu et al., 2021), Event-QA (Souza Costa et al., 2020), SimpleDBpediaQA (Azmy et al., 2018), CLC-QuAD (Zou et al., 2021), KQA Pro (Shi et al., 2020), SimpleQuestionsWikidata (Diefenbach et al., 2017), DBNQA (Yin et al., 2019), etc.

These datasets do not fulfill our current criteria and thus are not part of the initial version of the KGQA leaderboard. However, we encourage the community to help us update the leaderboard also for these datasets to prevent a replication crisis before it starts.

3.2. QA systems

While there are decentral collections of KGQA systems and there are available as code or Web service, e.g., https://github.com/semantic-systems/NLIWOD/tree/master/qa.systems, there is no up-to-date and systematically curated collection as of now. Our analysis shows that 24 provide a URL to a repository and 16 even to an online demo or Web API. However, after inspection, only 8 demos or Web APIs are still functional. This is the first hint toward an upcoming replication crisis. For a full list of systems, their descriptions, and pointers to their web services and demo, see `https://github.com/KGQA/leaderboard/blob/gh-pages/systems.md#Systems`

4. Dataset Analyses

We evaluated 100 papers and 98 systems focusing on 4 datasets, namely LC-QuAD version 1 and version 2 as well as the QALD-8 and 9 versions (all datasets released in 2017 or later). Figure 1 comprehensively summarizes the considered results of the leaderboard.

Based on the results, it became clear that *the evaluation values across the publications are often consistent*. The results contain multiple values for some of the system-dataset combinations (e.g., WDAqua-core0 over LC-QuAD 1.0), reported by different publications. Figure 2 demonstrates the evaluation values given a particular benchmark dataset grouped by the KGQA systems. For system-dataset combinations with multiple values, we calculated the standard deviation (std.). The std. values for such systems as QAKiS, TeBaQA, Elon, QASystem, gAnswer, and QAmp are not higher than 1%. This non-null std. is probably caused by the rounding errors. The only outliers in the evaluation values were observed given the WDAqua-core systems. For example, the paper (Zheng and Zhang, 2019) reports F1 Score of 38.7% for WDAqua-core0 over QALD-8, taking the results from the original publication of WDAqua-core0. Another paper (Orogat et al., 2021) reports F1 Score of only 33.0% for the same system-dataset combination. The authors (Orogat et al., 2021) calculated this result. The std. of both WDAqua-core versions reaches 9% on LC-QuAD 1.0 dataset and 3% on QALD-8. Note, the high std. values are not dependent on the datasets. Hence, the papers reporting significantly different results regarding WDAqua-core require further investigation. One of the assumptions is that WDAqua-core provides a publicly accessible demonstrator and

[6] https://www.dbpedia.org/

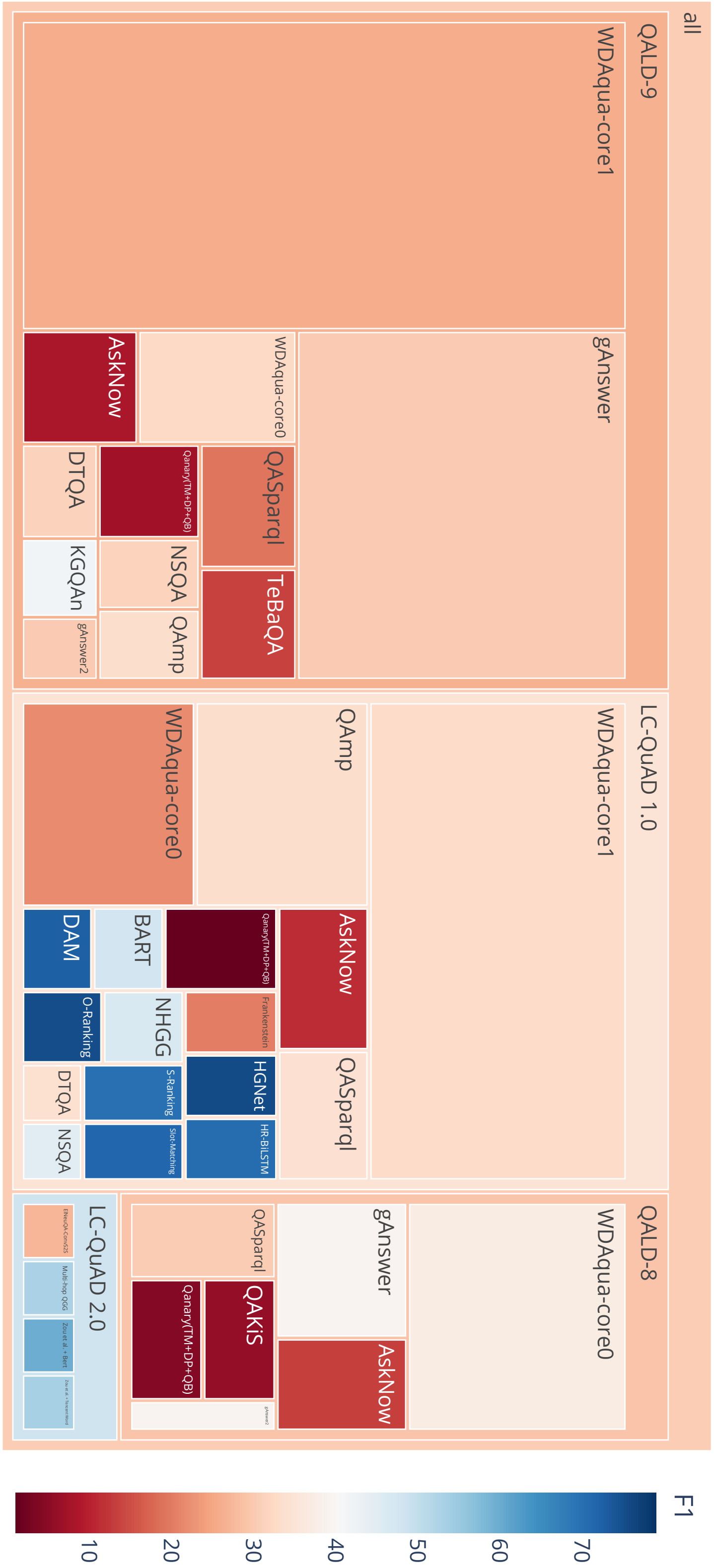


Figure 1: Treemap chart based on the collected results grouped by considered datasets (QALD-8, QALD-9, LC-QuAD 1.0, LC-QuAD 2.0). The KGQA systems are located within the dataset rectangles. The size of the rectangles is proportional to the number of mentions of a particular system in the whole leaderboard. The color of the rectangles denotes the average F1 Score of the corresponding systems. Only systems with more than 2 mentions are included.

API[7] which enables researchers to re-run the evaluation. This fact naturally implies possible differences in the evaluation results. However, there is no such systematic tendency for the other results as probably the majority of them were not reproduced but cited from the original publication. Despite the consistency of the results, the F1 Score values of the systems have a wide variance range given a particular dataset (cf., Figure 3). Surprisingly, the number of papers from ArXiv (preprints) in our leaderboard appears to be higher than the number of peer-reviewed papers (54% vs 46%). It was observed that the peer-reviewed papers report significantly higher results w.r.t. F1 Score which is 30.2% for preprints and 39.5% for peer-reviewed papers. The logical reason for this is that the peer-reviewed papers typically report state-of-the-art results, while preprints might contain preliminary work.

Given the considered results, it was observed that the authors of 72% papers did not include all the evaluation results from other publications in their comparison that were already available at a particular point in time. To find out this number, the set of systems from a publication reporting the values on particular datasets was compared to the set of systems released a year ago or earlier. For example, the publication (Orogat et al., 2021) released in 2021 does not consider the results of the QAmp system (Vakulenko et al., 2019) that was published in 2019.

5. Discussion

The trustworthiness of scientific results strongly depends on their comparability and replicability. In the field of KGQA, one could assume that the existence of a large and rising number QA datasets ensures comparability. Indeed, our analysis shows that the reported evaluations are *overwhelmingly coherent*. However, we observed several issues: first, the main reason why most numbers are identical is that people refer to results given in an original paper and its evaluation section. We could not find evidence that researchers actively tried to replicate results. A reason could be that only, 16 percent of the systems are available as source code (or web service/demo). However, even in the existence of an online demo, e.g., (Diefenbach et al., 2017; Diefenbach et al., 2020), the current state of the KGQA system seems not to be re-evaluated.

Second, our analysis indicates that researchers might have overlooked (best case) or omitted (worst case) relevant results that speak against their claims. For example, in (Wu et al., 2021) there are similar earlier works (Maheshwari et al., 2019; To and Reformat, 2020) which evaluated the same datasets and provided similar or even better results. However, we are well aware that researchers struggle with establishing an up-to-date overview of current research due to the time-consuming nature of the process without a central overview of KGQA systems.

Third, we see a strong need for improved evaluation methods. This demand can be covered by online evaluation methods, e.g., using platforms like Gerbil (Usbeck et al., 2019)). However, we also observed a decreasing amount of working online demos suggesting that a new form of a platform where models as such can be uploaded[8] could be a future direction.

Fourth, while developing new platforms and systems, we should also consider the rising critique on leaderboards regarding their utility for the NLP community at large (Ethayarajh and Jurafsky, 2020). Thus, we concur that evaluation protocols need to be published to foster transparency on leaderboards.

Finally, the lack of open-source implementations could be a starting point for a replication crisis. While there is no replication crisis in the field of KGQA as of now, the community needs to leverage novel initiatives such as the European Open Science Cloud[9] or the National Research Data Infrastructure for Data Science and AI[10]. Otherwise, models and source code might be lost or results will become incomparable in the long term.

6. Summary and Future Work

In this paper, we presented a novel community resource to track advances in the field of KGQA research. We foresee the need to maintain a KGQA focused platform as long as approaches such as ORKG (Auer et al., 2020) are not widely used or developed far enough. Of course, we could have just added our findings to reporting platforms. However, we believe, that this publication provides a more valuable base for discussions and reaches a wider audience than a silent upload. Additionally, since the QALD-9 evaluation campaign has passed for more than 3 years now, we intend to establish a central leaderboard to keep people on the same page.

In the future, we are looking into automatic ways to synchronize various reporting platforms with the KGQA leaderboard. We plan to extend the evaluation of QA systems, s.t., replicable evaluations, and data collections are possible. Additionally, improved metrics (e.g., (Orogat and El-Roby, 2021; Siciliani et al., 2021)) should be evaluated over models, source code, or via platforms to allow in-depth analyses of the capabilities of QA systems.

We are aware of research on other KGQA datasets grounded in Wikidata, Freebase, WikiMovies, and EventKG and want to encourage the community to update the KGQA leaderboard with the corresponding numbers.

7. Acknowledgements

The authors also acknowledge the financial support by the Federal Ministry for Economic Affairs and En-

[7] https://qanswer-frontend.univ-st-etienne.fr

[8] For example, https://project-hobbit.eu/outcomes/hobbit-platform/.

[9] https://eosc-portal.eu/

[10] https://www.nfdi4datascience.de/

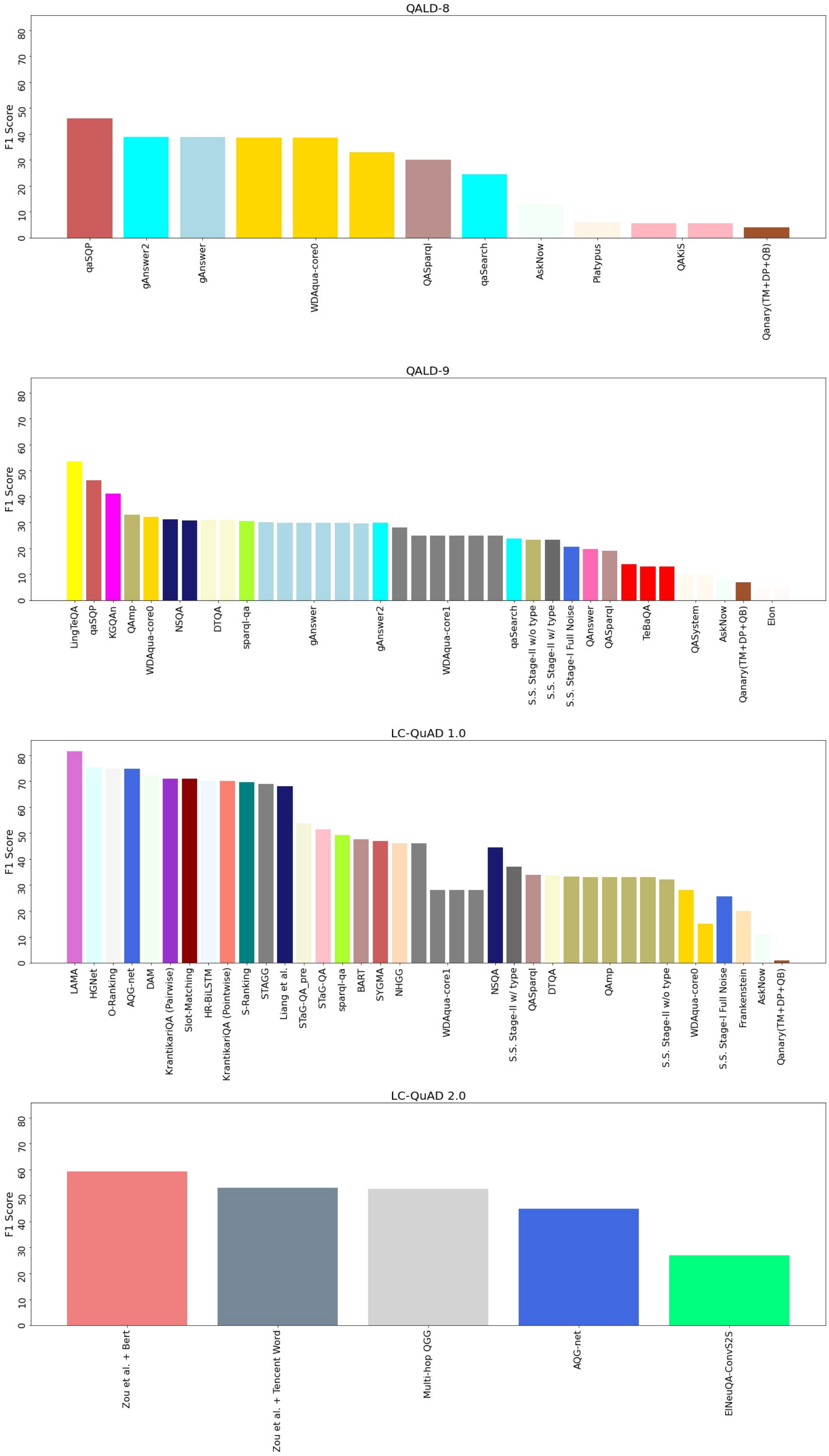


Figure 2: The chart demonstrates evaluation values (F1 Score) grouped by KGQA systems (same color) given a dataset. Each bar corresponds to a particular publication.

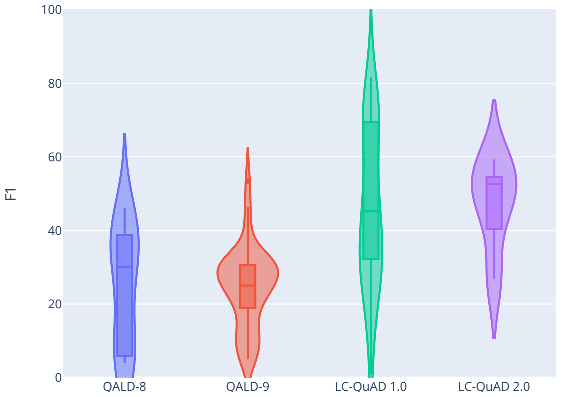


Figure 3: The figure demonstrates the distribution of the F1 Score values and their statistics from different publications given a dataset.

ergy of Germany in the project CoyPu (project number 01MK21007G).

8. Bibliographical References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K. E., Vogt, L., Prinz, M., Wiens, V., and Jaradeh, M. Y. (2020). Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis*, 44(3):516–529.

Both, A., Perevalov, A., Bartsch, J. R., Heinze, P., Iudin, R., Herkner, J. R., Schrader, T., Wunsch, J., Falkenhain, A. K., and Gürth, R. (2021). A question answering system for retrieving german COVID-19 data driven and quality-controlled by semantic technology. In Ilaria Tiddi, et al., editors, *Joint Proceedings of the Semantics co-located events: Poster&Demo track and Workshop on Ontology-Driven Conceptual Modelling of Digital Twins co-located with Semantics 2021, Amsterdam and Online, September 6-9, 2021*, volume 2941 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Chakraborty, N., Lukovnikov, D., Maheshwari, G., Trivedi, P., Lehmann, J., and Fischer, A. (2021). Introduction to neural network-based question answering over knowledge graphs. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 11(3).

Chandra, A., Fahrizain, A., Laufried, S. W., et al. (2021). A survey on non-english question answering dataset. *arXiv preprint arXiv:2112.13634*.

Diefenbach, D., Singh, K., and Maret, P. (2017). Wdaqua-core0: A question answering component for the research community. In Mauro Dragoni, et al., editors, *Semantic Web Challenges*, pages 84–89, Cham. Springer International Publishing.

Diefenbach, D., López, V., Singh, K. D., and Maret, P. (2018). Core techniques of question answering systems over knowledge bases: a survey. *Knowl. Inf. Syst.*, 55(3):529–569.

Diefenbach, D., Both, A., Singh, K., and Maret, P. (2020). Towards a question answering system over the semantic web. *Semantic Web*, 11(3):421–439.

Diefenbach, D., Wilde, M. D., and Alipio, S. (2021). Wikibase as an infrastructure for knowledge graphs: The eu knowledge graph. In *International Semantic Web Conference*, pages 631–647. Springer.

Dubey, M., Banerjee, D., Chaudhuri, D., and Lehmann, J. (2018). EARL: joint entity and relation linking for question answering over knowledge graphs. In Denny Vrandecic, et al., editors, *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I*, volume 11136 of *Lecture Notes in Computer Science*, pages 108–126. Springer.

Dubey, M., Banerjee, D., Abdelkawi, A., and Lehmann, J. (2019). Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In Chiara Ghidini, et al., editors, *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, volume 11779 of *Lecture Notes in Computer Science*, pages 69–78. Springer.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., and Vrandečić, D. (2014). Introducing wikidata to the linked data web. In *International semantic web conference*, pages 50–65. Springer.

Ethayarajh, K. and Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. In Bonnie Webber, et al., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4846–4853. Association for Computational Linguistics.

Fu, B., Qiu, Y., Tang, C., Li, Y., Yu, H., and Sun, J. (2020). A survey on complex question answering over knowledge base: Recent advances and challenges. *CoRR*, abs/2007.13069.

Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., and Ngomo, A. N. (2017). Survey on challenges of question answering in the semantic web. *Semantic Web*, 8(6):895–920.

Hu, S., Zou, L., Yu, J. X., Wang, H., and Zhao, D. (2018). Answering natural language questions by subgraph matching over knowledge graphs (extended abstract). In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018*, pages 1815–1816. IEEE Computer Society.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. (2020). Open graph benchmark: Datasets for machine learning on graphs. In Hugo Larochelle, et al., editors, *Advances* *in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jurafsky, D. and Martin, J. H. (2018). Speech and language processing (draft). *preparation [cited 2022 January 4] Available from: https://web. stanford. edu/˜jurafsky/slp3*.

Maheshwari, G., Trivedi, P., Lukovnikov, D., Chakraborty, N., Fischer, A., and Lehmann, J. (2019). Learning to rank query graphs for complex question answering over knowledge graphs. In Chiara Ghidini, et al., editors, *The Semantic Web – ISWC 2019*, pages 487–504, Cham. Springer International Publishing.

Mutabazi, E., Ni, J., Tang, G., and Cao, W. (2021). A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12).

Orogat, A. and El-Roby, A. (2021). Cbench: Demonstrating comprehensive evaluation of question answering systems over knowledge graphs through deep analysis of benchmarks. *Proc. VLDB Endow.*, 14(12):2711–2714.

Orogat, A., Liu, I.-C., and El-Roby, A. (2021). Cbench: Towards better evaluation of question answering over knowledge graphs. *Proc. VLDB Endow.*, 14:1325–1337.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, et al., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Saleem, M., Dastjerdi, S. N., Usbeck, R., and Ngomo, A.-C. N. (2017). Question answering over linked data: What is difficult to answer? what affects the f scores? In *BLINK/NLIWoD3@ ISWC*.

Siciliani, L., Basile, P., Lops, P., and Semeraro, G. (2021). Mqald: Evaluating the impact of modifiers in question answering over knowledge graphs. *Semantic Web*.

Singh, K., Radhakrishna, A. S., Both, A., Shekarpour, S., Lytra, I., Usbeck, R., Vyas, A., Khikmatullaev, A., Punjani, D., Lange, C., Vidal, M., Lehmann, J., and Auer, S. (2018). Why reinvent the wheel: Let's build question answering systems together. In Pierre-Antoine Champin, et al., editors, *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1247–1256. ACM.

Singh, K., Saleem, M., Nadgeri, A., Conrads, F., Pan, J. Z., Ngomo, A. N., and Lehmann, J. (2019). Qaldgen: Towards microbenchmarking of question answering systems over knowledge graphs. In Chiara Ghidini, et al., editors, *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II*, volume 11779 of *Lecture Notes in Computer Science*, pages 277–292. Springer.

To, N. D. and Reformat, M. (2020). Question-answering system with linguistic terms over rdf knowledge graphs. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4236–4243.

Trivedi, P., Maheshwari, G., Dubey, M., and Lehmann, J. (2017). Lc-quad: A corpus for complex question answering over knowledge graphs. In Claudia d'Amato, et al., editors, *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II*, volume 10588 of *Lecture Notes in Computer Science*, pages 210–218. Springer.

Usbeck, R., Gusmita, R. H., Ngomo, A. N., and Saleem, M. (2018). 9th challenge on question answering over linked data (QALD-9) (invited paper). In Key-Sun Choi, et al., editors, *Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, California, United States of America, October 8th - 9th, 2018*, volume 2241 of *CEUR Workshop Proceedings*, pages 58–64. CEUR-WS.org.

Vakulenko, S., Fernandez Garcia, J. D., Polleres, A., de Rijke, M., and Cochez, M. (2019). Message passing for complex question answering over knowledge graphs. In *Proceedings of the 28th acm international conference on information and knowledge management*, pages 1431–1440.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Wu, P., Wu, Y., Wu, L., Zhang, X., and Feng, Z. (2021). Modeling global semantics for question answering over knowledge bases.

Zheng, W. and Zhang, M. (2019). Question answering over knowledge graphs via structural query patterns. *ArXiv*, abs/1910.09760.

9. Language Resource References

Azmy, M., Shi, P., Lin, J., and Ilyas, I. (2018). Farewell freebase: Migrating the simplequestions dataset to dbpedia. In *Proceedings of the 27th international conference on computational linguistics*, pages 2093–2103.

Bao, J., Duan, N., Yan, Z., Zhou, M., and Zhao, T. (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.

Berant, J., Chou, A., Frostig, R., and Liang, P. (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.

Bordes, A., Usunier, N., Chopra, S., and Weston, J. (2015). Large-scale simple question answering with memory networks. *CoRR*, abs/1506.02075.

Cai, Q. and Yates, A. (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.

Cui, R., Aralikatte, R., Lent, H., and Hershcovich, D. (2021). Multilingual compositional wikidata questions. *arXiv preprint arXiv:2108.03509*.

Diefenbach, D., Tanon, T. P., Singh, K. D., and Maret, P. (2017). 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*.

Gu, Y., Kase, S., Vanni, M., Sadler, B., Liang, P., Yan, X., and Su, Y. (2021). Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Jia, Z., Abujabal, A., Saha Roy, R., Strötgen, J., and Weikum, G. (2018). Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062.

Jia, Z., Pramanik, S., Saha Roy, R., and Weikum, G. (2021). Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 792–802.

Jiang, K., Wu, D., and Jiang, H. (2019). Freebaseqa: a new factoid qa data set matching trivia-style question-answer pairs with freebase. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 318–323.

Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., et al. (2019). Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*.

Korablinov, V. and Braslavski, P. (2020). Rubq: a russian dataset for question answering over wikidata. In *International Semantic Web Conference*, pages 97–110. Springer.

Perevalov, Aleksandr and Diefenbach, Dennis and Usbeck, Ricardo and Both, Andreas. (2022). *QALD-9-plus: A Multilingual Dataset for Question Answering over DBpedia and Wikidata Translated by Native Speakers*.

Rybin, I., Korablinov, V., Efimov, P., and Braslavski, P. (2021). Rubq 2.0: An innovated russian question answering dataset. In *European Semantic Web Conference*, pages 532–547. Springer.

Saxena, A., Chakrabarti, S., and Talukdar, P. (2021). Question answering over temporal knowledge graphs. *arXiv preprint arXiv:2106.01515*.

Serban, I. V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., and Bengio, Y. (2016). Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*.

Shi, J., Cao, S., Pan, L., Xiang, Y., Hou, L., Li, J., Zhang, H., and He, B. (2020). Kqa pro: A large diagnostic dataset for complex question answering over knowledge base. *arXiv e-prints*, pages arXiv–2007.

Souza Costa, T., Gottschalk, S., and Demidova, E. (2020). Event-qa: A dataset for event-centric question answering over knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3157–3164.

Su, Y., Sun, H., Sadler, B., Srivatsa, M., Gür, I., Yan, Z., and Yan, X. (2016). On generating characteristic-rich question sets for qa evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 562–572.

Talmor, A. and Berant, J. (2018). The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Usbeck, R., Röder, M., Hoffmann, M., Conrads, F., Huthmann, J., Ngomo, A. N., Demmler, C., and Unger, C. (2019). Benchmarking question answering systems. *Semantic Web*, 10(2):293–304.

Yih, W.-t., Richardson, M., Meek, C., Chang, M.-W., and Suh, J. (2016). The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206.

Yin, X., Gromann, D., and Rudolph, S. (2019). Neural machine translating from natural language to sparql. *arXiv preprint arXiv:1906.09302*.

Zhang, Y., Dai, H., Kozareva, Z., Smola, A., and Song, L. (2018). Variational reasoning for question answering with knowledge graph. In *AAAI*.

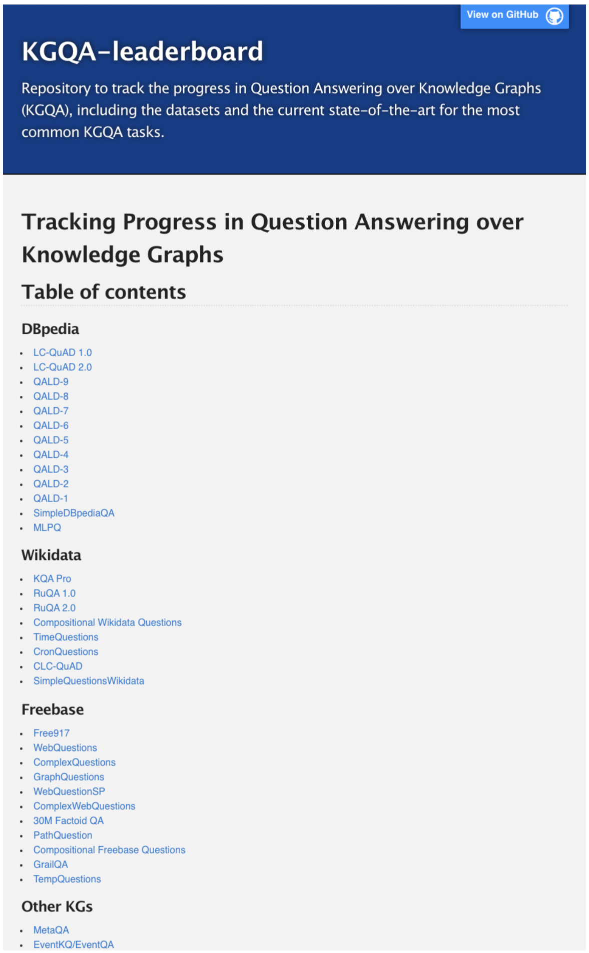
Zhou, M., Huang, M., and Zhu, X. (2018). An interpretable reasoning network for multi-relation question answering. *arXiv preprint arXiv:1801.04726*.

Zou, J., Yang, M., Zhang, L., Xu, Y., Pan, Q., Jiang, F., Qin, R., Wang, S., He, Y., Huang, S., et al. (2021). A chinese multi-type complex questions answering dataset over wikidata. *arXiv preprint arXiv:2111.06086*.

A. KGQA Leaderboard

To ensure the replication of KGQA systems and the trustworthiness of their evaluation results, we provide a leaderboard. The leaderboard is available at https://kgqa.github.io/leaderboard/. It can be used to compare the capabilities of these KGQA systems over the latest and commonly used KGQA benchmark datasets by tracking the progress. It includes the datasets, links, papers, and SOTA results.

At the time of writing, the leaderboard includes a total of 34 KGQA datasets across 5 knowledge graphs (i.e., DBpedia, Wikidata, Freebase, WikiMovies, and EventKG). As shown in Fig. 4, these KGQA datasets are separated by the used target KGs. Fig. 5 shows an example of LCQuAD V1.0 Leaderboard. We will continuously add newly released datasets and their SOTA results, and invite other researchers to make their contributions by adding new results based on these KGQA dataset overviews.



KGQA-leaderboard

Repository to track the progress in Question Answering over Knowledge Graphs (KGQA), including the datasets and the current state-of-the-art for the most common KGQA tasks.

Tracking Progress in Question Answering over Knowledge Graphs

Table of contents

DBpedia

- LC-QuAD 1.0
- LC-QuAD 2.0
- QALD-9
- QALD-8
- QALD-7
- QALD-6
- QALD-5
- QALD-4
- QALD-3
- QALD-2
- QALD-1
- SimpleDBpediaQA
- MLPQ

Wikidata

- KQA Pro
- RuQA 1.0
- RuQA 2.0
- Compositional Wikidata Questions
- TimeQuestions
- CronQuestions
- CLC-QuAD
- SimpleQuestionsWikidata

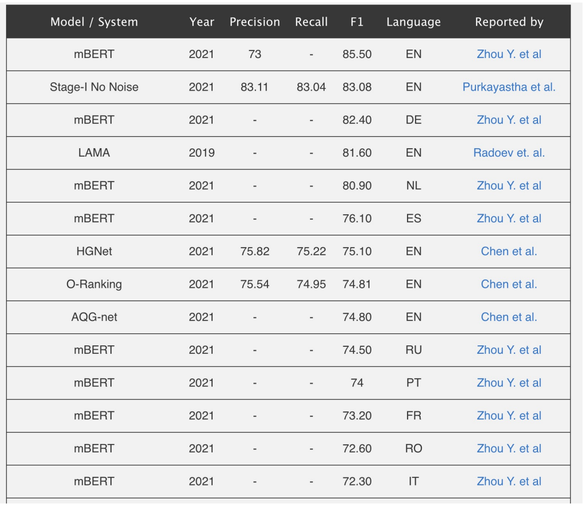
Freebase

- Free917
- WebQuestions
- ComplexQuestions
- GraphQuestions
- WebQuestionSP
- ComplexWebQuestions
- 30M Factoid QA
- PathQuestion
- Compositional Freebase Questions
- GrailQA
- TempQuestions

Other KGs

- MetaQA
- EventKQ/EventQA

Figure 4: Interface of the KGQA leaderboard.



Model / System	Year	Precision	Recall	F1	Language	Reported by
mBERT	2021	73	-	85.50	EN	Zhou Y. et al
Stage-I No Noise	2021	83.11	83.04	83.08	EN	Purkayastha et al.
mBERT	2021	-	-	82.40	DE	Zhou Y. et al
LAMA	2019	-	-	81.60	EN	Radoev et. al.
mBERT	2021	-	-	80.90	NL	Zhou Y. et al
mBERT	2021	-	-	76.10	ES	Zhou Y. et al
HGNet	2021	75.82	75.22	75.10	EN	Chen et al.
O-Ranking	2021	75.54	74.95	74.81	EN	Chen et al.
AQG-net	2021	-	-	74.80	EN	Chen et al.
mBERT	2021	-	-	74.50	RU	Zhou Y. et al
mBERT	2021	-	-	74	PT	Zhou Y. et al
mBERT	2021	-	-	73.20	FR	Zhou Y. et al
mBERT	2021	-	-	72.60	RO	Zhou Y. et al
mBERT	2021	-	-	72.30	IT	Zhou Y. et al

Figure 5: An example of LC-QuAD V1.0 Leaderboard.