rewordQALD9: A Bilingual Benchmark with Alternative Rewordings of QALD Questions

Manuela Sanguinetti[1,*], Maurizio Atzori[1] and Nicoletta Puddu[2]

[1]*Department of Mathematics and Computer Science, University of Cagliari, Italy*
[2]*Dipartimento di Lettere, Lingue e Beni Culturali, University of Cagliari, Italy*

Abstract

In this paper we describe an extended version of the QALD dataset, a well-known benchmark resource used for the task of Question Answering over knowledge graphs (KGQA). Along the lines of similar projects, the purpose of this work is to make available a) high-quality data even for languages other than English, and b) multiple reformulations of the same question, to test systems' robustness. The QALD version we used is the one released for the 9th edition of the challenge of Question Answering over Linked Data, and the languages involved are English and Italian. Besides a revised and improved quality of Italian translations of questions, the resource presented here features a number of alternative rewordings of both English and Italian questions. The usability of the resource has been tested on the QAnswer multilingual Question Answering system and through the GERBIL platform. The resource has been publicly released for research purposes.

Keywords

question answering, knowledge graphs, benchmark, paraphrases

1. Introduction

The task of Question Answering over Knowledge Graphs (KGQA) consists precisely in mapping a natural language question into a formal query - most typically in SPARQL - to an underlying knowledge graph, such as Wikidata, DBpedia or Freebase. Regardless of the approach adopted (whether based on templates, semantic parsing, or end-to-end neural approaches), the availability of data for both training and testing systems is critical.

A well-known benchmark dataset for this task is the one released in conjunction with the challenge series on Question Answering over Linked Data (QALD) [1], just recently in its tenth edition.[1] It consists of a set of questions accompanied by the corresponding SPARQL query, along with the answer (or list of answers) and the information on the answer type. Except for the latest edition of the QALD challenge, where Wikidata has been adopted as underlying KG, questions and related queries in the previous dataset versions used to target DBpedia. Besides representing a widely used dataset for benchmarking, the QALD dataset has also served multiple purposes, among which the automatic extraction of query templates [2, 3], the development

SEMANTICS 2022 EU: 18th International Conference on Semantic Systems, September 13-15, 2022, Vienna, Austria
*Corresponding author.
manuela.sanguinetti@unica.it (M. Sanguinetti); atzori@unica.it (M. Atzori); n.puddu@unica.it (N. Puddu)
0000-0002-0147-2208 (M. Sanguinetti); 0000-0001-6112-7310 (M. Atzori); 0000-0003-4924-0825 (N. Puddu)

[1]https://www.nliwod.org/challenge

of micro-benchmarking platforms [4] or the creation of QALD-like resources based on its extension [5, 6]. Another aspect that characterizes the resource since QALD3 – and which still constitutes one of the main challenges in KGQA – is multilinguality: the same question is in fact expressed in multiple languages (up to 11 in QALD9), thus enabling the development of KGQA systems that also support languages other than English. However, the quality of non-English translations in the dataset is sometimes poor, with sentences that are unnatural, if not entirely ungrammatical. A recent paper addressed these issues [5] by working specifically on improving the quality of translated questions for a sub-set of QALD9 languages (French, German, and Russian) and adding brand new translations into underrepresented languages, such as Armenian, Belarusian, Lithuanian, Bashkir, and Ukrainian.

The work introduced here partially follows a similar path, in that it aims to provide manually-curated translations of Italian questions from the QALD9 dataset. Among the main goals of our work is also the development of a resource that could be used as a testbed for systems' robustness and flexibility in handling a wider range of language variations, possibly including both standard and informal register. Therefore, a further enhancement of the resource introduced with this work is the addition of question reformulations to the original QALD9 data. Besides the increased size of the final resource, we believe the main contributions of this work are 1) the improved quality of questions for Italian, 2) the availability of alternative rewordings of the same questions, that could be useful as adversarial examples (similar in spirit to Ribeiro et al. [7]) or for the development of template-based KGQA systems.

The rest of the paper will provide a brief description of the resulting dataset, that we called *rewordQALD9*, and a preliminary baseline evaluation obtained with QAnswer [8], a state-of-the-art multilingual KGQA system. The resource is freely available at the following link: https://github.com/msang/rewordQALD9

2. Dataset Development and Description

The two main tasks required for the development of *rewordQALD9* were translation revision and question reformulation that possibly include informal registers; for this purpose, eight undergraduate students majoring in translation studies were involved in the project. They were all Italian native speakers proficient in English. As for the second task in particular, specific guidelines were provided to the annotators, according to which question reformulations should consist of more or less pervasive changes in the sentence structure or lexicon; we recommended to alter the question form as much as possible, as long as its meaning is still preserved. Proposed changes included (but were not limited to) use of different parts of speech, diathesis alternation and different verb tenses, and use of synonyms. To further enhance variety in language forms, paraphrases were expected to include even linguistic phenomena more typical of informal registers (e.g. dislocations, clefts, etc.) and features that mimic spoken language (as if the question were posed to a conversational agent/personal assistant).

Once these tasks were completed, a pairwise review was carried out, so that each annotator had the opportunity to verify another annotator's work. Finally, a third annotator, not involved in the previous stages, performed a final check over the entire dataset, to ensure that both revisions and paraphrases were consistent with the task specifications. In addition, duplicate or

Table 1
Overview of the final dataset.

Lang.	Split	Orig. questions	Paraphrases	Tot. Questions	
En	Train	405	729	1134	1546
	Test	146	266	412	
It	Train	405	843	1248	1707
	Test	146	313	459	

near-duplicate questions in the original dataset were removed (for example, the question "Where did Abraham Lincoln die" appeared both in the training and test set). Annotators were able to provide an average of two paraphrases per question (on a range of one to three questions), with a higher proportion of paraphrases in the Italian section. The resulting dataset thus consists of 995 additional questions for English and 1,156 for Italian, and a total amount of around 1.5K and 1.7K questions for English and Italian, respectively, as also summarized in Table 1.[2]

To ensure its usability and compatibility with well-known benchmarking platforms such as GERBIL [9], the dataset has been encoded in the QALD-JSON format.[3]

Before testing the resource on available KGQA systems, we first carried out some preliminary qualitative analysis with the aim to provide an overall picture of the data at hand by means of simple descriptive measures. We considered in particular five aspects for the analysis, all reflecting the possible challenging nature of the additional questions to the QA task: they range from **average question length** and **lexical diversity** to **syntactic** and **semantic similarity**. The average question length was measured both in terms of *average character count* and *average number of tokens* per question, while lexical diversity is expressed through the *type/token ratio* (i.e. the number of unique tokens over the total number of tokens in a question). All three were computed using the LinguaF library,[4] as in Perevalov et al. [5]. *Syntactic and semantic similarities* were taken into account in this analysis to verify whether annotators effectively reworded original questions using structurally different forms, while still conveying the same content. Therefore, comparisons were made between each pair of original and paraphrased questions. To compute syntactic similarity, all questions were first parsed using the UDPipe REST service,[5] [10] and the pairs of syntactic trees were then compared using the `block.eval.f1` module of UDAPI[6] [11]. The resulting $F1$ score has been used as a proxy measure to determine the structural similarity between the tree pairs. Finally, semantic similarity was computed using a multilingual pre-trained language model based on Sentence-BERT [12][7] and cosine similarity as metric. All the results obtained with these measures were averaged over the whole dataset,

[2]In terms of time, it took around 25h for each annotator to process an average of 69 English-Italian question pairs. Also, an intermediate and a final meeting were scheduled to discuss particularly difficult or possibly ambiguous questions.
[3]https://github.com/dice-group/gerbil/wiki/Question-Answering#web-service-interface
[4]https://github.com/Perevalov/LinguaF
[5]https://lindat.mff.cuni.cz/services/udpipe/
[6]https://udapi.readthedocs.io/en/latest/udapi.block.eval.html#module-udapi.block.eval.f1
[7]Hosted on the HuggingFace Model Hub: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Table 2
Dataset basic statistics.

	EN	IT
Avg. question length (tok./sent.)	7.69	7.96
Avg. question length (char./sent.)	35.84	39.38
Lexical diversity (type/tok.)	0.18	0.19
Syntactic similarity (F1)	41.32%	51.21%
Semantic similarity ($cos(\theta)$)	0.91	0.92

Table 3
Baseline evaluation on QAnswer (metric: F1-QALD).

	EN	IT
QALD9	0.3235	0.2826
rewordQALD9	0.3186	0.2603

as shown in Table 2. What reported in the table shows in particular that lexical diversity is quite low, meaning that annotators tended to use the same terms as the ones in the original questions, but resorting to alternative structural forms that overall did not alter the meaning of the original question. While a higher lexical variability would have been preferable, the latter two measures show the suitability of the resource to test systems robustness to alternative reformulations of a same question.

3. Baseline Evaluation and Future Work

To test the resource usability, as well as the possible impact of the reworded questions on the QA task, we compared the results of a multilingual KGQA system on the original QALD9 questions and on the *rewordQALD9* introduced here. For our experiments, we selected QAnswer [8] as it is, to the best of our knowledge, the only available system that supports both English and Italian (among other languages). System's performance was evaluated on GERBIL[8] and according to the F1-QALD metric, as computed in the platform. Results are reported in Table 3.

As expected, performance consistently decreases in *rewordQALD9*, mainly due to the increased difficulty the system has to face in handling reformulations: as a matter of fact, a separate evaluation on the group of paraphrases only reported a F1-QALD = 0.3113 for English and 0.244 for Italian, thus confirming the more challenging nature of reworded questions for the QA system.

The baseline evaluation just described, although not expanded here due to space constraints, paves the way to further investigations that assess the resource usefulness for micro-benchmarking purposes (along the lines of Singh et al. [4]), e.g. measuring the impact of given reformulations on the system's output accuracy. As future work, we intend to provide a more systematic account of such impact, both comparing multiple KGQA systems (whenever at least one of the two dataset languages are supported) and verifying possible correlations between the systems' output and the descriptive measures defined in Section 2. Furthermore, although conducted independently, this work has several aspects in common with what proposed in Perevalov et al. [5], especially regarding motivations and approach. The eventual integration of the two projects is certainly a hopeful outcome, with a view to further extending and improving the original QALD benchmark in terms of both multilinguality and interoperability among resources.

[8] www.gerbil-qa.aksw.org/gerbil/config

Acknowledgments

We warmly thank the students who participated in this work: V. Carta, F. Carrucciu, M. Desogus, S. Enis, G. Fanari, E. Massa, M. Montis, and C. Sannia. We also thank Aleksandr Perevalov for his help in using QAnswer. The work of M. Sanguinetti and M. Atzori is partially funded by PRIN 2017 (2019-2022) project *HOPE* - High quality Open data Publishing and Enrichment and Fondazione di Sardegna project *ASTRID* - Advanced learning STRategies for high-dimensional and Imbalanced Data (CUP F75F21001220007).

References

[1] R. Usbeck, R. Hari Gusmita, M. Saleem, A.-C. Ngonga Ngomo, 9th Challenge on Question Answering over Linked Data (QALD-9), in: Joint proceedings of SemDeep4 and NLIWoD4, CEUR-WS.org, 2018.

[2] A.-K. Hartmann, E. Marx, T. Soru, Generating a large dataset for neural question answering over the DBpedia knowledge base, in: Workshop on Linked Data Management, 2018.

[3] D. Vollmers, R. Jalota, D. Moussallem, H. Topiwala, A. N. Ngomo, R. Usbeck, Knowledge graph question answering using graph-pattern isomorphism, CoRR (2021). URL: https://arxiv.org/abs/2103.06752.

[4] K. Singh, M. Saleem, A. Nadgeri, F. Conrads, J. Z. Pan, A.-C. N. Ngomo, J. Lehmann, Qaldgen: Towards microbenchmarking of question answering systems over knowledge graphs, in: C. e. a. Ghidini (Ed.), The Semantic Web – ISWC 2019. Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 277–292.

[5] A. Perevalov, D. Diefenbach, R. Usbeck, A. Both, Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers, in: Proceedings of ICSC, 2022, pp. 229–234.

[6] L. Siciliani, P. Basile, P. Lops, G. Semeraro, MQALD: Evaluating the impact of modifiers in Question Answering over Knowledge Graphs, Semantic Web 1 (2021) 1–14.

[7] M. T. Ribeiro, S. Singh, C. Guestrin, Semantically equivalent adversarial rules for debugging NLP models, in: Proceedings of ACL, Association for Computational Linguistics, 2018, pp. 856–865.

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

[9] R. Usbeck, M. Röder, M. Hoffmann, F. Conrads, J. Huthmann, A. Ngonga Ngomo, C. Demmler, C. Unger, Benchmarking question answering systems, Semantic Web 10 (2019) 293–304.

[10] M. Straka, UDPipe 2.0 prototype at CoNLL 2018 UD shared task, in: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2018, pp. 197–207.

[11] M. Popel, Z. Zabokrtský, M. Vojtek, Udapi: Universal API for Universal Dependencies, in: M. de Marneffe, J. Nivre, S. Schuster (Eds.), Proceedings of the NoDaLiDa Workshop on Universal Dependencies, Association for Computational Linguistics, 2017, pp. 96–101.

[12] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of EMNLP, Association for Computational Linguistics, 2019.