When context matters: Entity Linking in the scholarly domain

Nadine Steinmetz

University of Applied Sciences Erfurt, Germany

Abstract

In the field of Question Answering on Knowledge Graphs (KGQA), entity linking is an essential substep to transform a natural language (NL) question to the formal query language SPARQL. The required entities need to be spotted in the text and the correct resource in the knowledge graph (KG) has to be identified. The latter step is especially hard for entities in questions when there is only little or no context. We already presented a previous approach using abstract meaning representation (AMR) of the question to spot the surface forms of entities. With this paper, we present our adapted approach for the domain of scholarly questions – in specific DBLP QuAD.

Keywords

entity linking, kgqa, dblp quad

1. Introduction

Entity linking is an essential part of the semantic analysis of natural language text, but specifically important to transform NL questions to a formal query language in the field of Question Answering on Knowledge Graphs (KGQA).

Considering the sample question *Who painted The Storm on the Sea of Galilee?*, the surface form of the named entity can consist of multiple different words and types of words. For our example, the surface form of the named entity `The Storm on the Sea of Galilei` consists of seven words and three different word types: determiners, prepositions and nouns. The correct identification of the surface form is even harder if the question is case insensitive.

The disambiguation of surface forms to the correct named entity often requires the question answering (QA) system to choose between several possible entities. For KGQA usually the underlying knowledge base holds context information for each alternative which must be compared to the context information of the input question. But, in QA the input context is often very little respectively not existing.

We present our adapted approach on entity linking in the context of KGQA. The original approach has been evaluated on QALD-9 and LC-QuAD 2.0 – which are based on all-purpose KGs DBpedia and Wikidata [1]. With this paper, we examine required adaptations of the

Scholarly QALD @ ISWC 2023: Subtasks for Question Answering over Scholarly Knowledge Graphs , November 6 - 11, 2023, Athens, Greece

nadine.steinmetz@fh-erfurt.de (N. Steinmetz)

http://www.fh-erfurt.de/steinmetz (N. Steinmetz)

0000-0003-3601-7579 (N. Steinmetz)

approach when the underlying KG is modelling a very specific domain, such as the DBLP RDF KG.

The general approach utilizes the abstract meaning representation (AMR) of the question. Within the syntactic structure of the graph, named entities are represented as name nodes and the surface form is included as child nodes. For the subsequent entity linking process, a extensive entity dictionary is required as well as context information for all entities and from the input question. We examined several options for the DBLP RDF KG and the DBLP-QuAD.

The remainder of the paper is structured as follows: Related work is described in Section 2. The entity linking approach including the description of the dictionary and ranking options is depicted in Section 3. With our approach, we took part in the Scholarly QALD challenge and the results are discussed in Section 4. We summarize our approach and discuss future work in Section 5.

2. Related Work

One of the first entity linking approaches has been introduced by Milne et al. in 2008 [2]. In the following years, more approaches were introduced such as DBpedia Spotlight[3] or tagme[4]. These approaches can handle longer texts and therefore utilize a reasonable amount of context information. In contrast, Falcon 2.0 has been developed by Sakor et al. for short texts respectively a single sentence or a question [5]. The authors provide an API which enables users to detect named entities from Wikidata and DBpedia.

As we are using the AMR graph for the generation of the SPARQL query, other approaches based on AMR graphs might be interesting. Besides our own approach [6], there are already other existing approaches on KGQA utilizing AMR in the transformation pipeline from NL to SPARQL, as e.g. [7] and [8].

Both latter approaches do not use the AMR graph itself, but BLINK[9] for the entity linking process.

3. Approach

Our presented approach is based on the entity linking process introduced in [1]. In order to generate a SPARQL query from the natural language question, we also deduce the mentioned named entities from the AMR graph. As described in [1], we retrained the AMR model with augmented training data in order to improve the quality of the AMR generation also for questions with incorrect casing. Named entities are identified in the graph as name nodes and surface forms are child nodes of the name node. The surface forms are used for a lookup in our entity dictionary. The dictionary contains main and alternative labels of all entities of the respective knowledge graphs. In addition, the dictionary contains contextual information about the entities:

- indegree (as popularity measure)
- similarity score of label to main label (for alternative labels)
- contextual descriptions (abstracts for Wikidata and DBpedia entities)

Using this additional information, scores are calculated for the entities and they are disambiguated in case of ambiguous surface forms. These disambiguation scores are further referenced as popularity score, label score and context score. With this approach, we achieved very good results for the QALD 9.0 and LC-QuAD 2.0 datasets. For the DBLP QALD dataset, the approach had to adapted in several aspects:

- pre-processing of NL questions for paper titles and author names
- generation of entity dictionary with alternative labels and context information from DBLP RDF
- exclusion of publishers as named entities
- disambiguation using complete question

We will further describe the specific approach for the DBLP QALD questions in the following sections.

3.1. AMR generation

The DBLP QUAD dataset contains questions about authors and publications. Paper titles are enclosed by quotation marks which, in general, makes it easy to identity the surface form of the titles. As we utilize the AMR graph for the further generation of the SPARQL query, we aim to maintain the linguistic embedding of the entities within the NL question and therefore focus on identifying the named entities (including paper titles) within the AMR graph. Unfortunately, AMR graphs and their generation have some limitations, including that it "drops grammatical number, tense, aspect, quotation marks, etc."[1]. Dropping the quotation marks results in embedding the paper title in the context of the question and using parts of title in the linguistic structure. For instance, the question *How many authors does 'Measuring the impact of temporal context on video retrieval' have?* results in an incorrect AMR graph as shown of the left below. The correct graph is shown on the right:

Incorrect AMG graph
with dropped quotation marks:

```
(h / have-03
      :ARG0 (p / publication
            :name (n / name
                  :op1 "Measuring"
                  :op2 "The"
                  :op3 "Impact"
                  :op4 "of"
                  :op5 "Time"
                  :op6 "Context"))
      :ARG1 (p2 / person
            :ARG0-of (a / author-01)
            :quant (a2 / amr-unknown))
      :ARG2 (r / retrieve-01
            :ARG1 (v / video)))
```

Correct AMR graph:

```
(h / have-03
      :ARG0 (p / publication
            :name (n / name
                  :op1 "Measuring"
                  :op2 "The"
                  :op3 "Impact"
                  :op4 "of"
                  :op5 "Temporal"
                  :op6 "Context"
                  :op7 "On"
                  :op8 "Video"
                  :op9 "Retrieval"))
      :ARG1 (p2 / person
            :ARG0-of (a / author-01)
            :quant (a2 / amr-unknown))
```

[1] https://github.com/amrisi/amr-guidelines/blob/master/amr.md

Obviously, there are two problems with the handling of the question:

1. the surface forms of the paper titles are changed (temporal → time)
2. parts of the paper title are excluded from the name node (video retrieval as separate node under `:ARG2`)

We therefore use a placeholder for the paper titles before the AMR generation and then replace the placeholder with the actual title at the correct position within the AMR graph. The respective AMR graph for the question including the placeholder looks like this:

```
(h / have-03
      :ARG0 (p / publication
            :name (n / name
                  :op1 "PaperOne"))
      :ARG1 (p2 / person
            :ARG0-of (a / author-01)
            :quant (a2 / amr-unknown))
```

In this way, we preserve the syntactic structure of the question and the position of the paper title within the graph.

Another issue with the generation of the AMR graphs which seems to be specific to the DBLP QuAD dataset, is the mentioning of author names in certain ways. In some cases, authors are mentioned in the question with their lastname first following the given name separated by a comma, as e.g. in *What are the papers written by Ben-Simon, E. and Andrey Zhdanov together?*. The AMR for that question looks like this:

```
(p / paper
      :ARG1-of (w / write-01
            :ARG0 (a / and
                  :op1 (p2 / person
                        :name (n / name
                              :op1 "Ben-Simon"))
                  :op2 (p3 / person
                        :name (n2 / name
                              :op1 "E."))
                  :op3 (p4 / person
                        :name (n3 / name
                              :op1 "Andrey"
                              :op2 "Zhdanov"))
                  :mod (t / together)))
      :domain (a2 / amr-unknown))
```

Apparently, the comma is treated like a separator for a list of person names. We therefore preprocess the question using regular expressions to identify reverse author names and rewrite the author names in the pattern `[first name(s)][space][last name(s)]` before generating the AMR graph.

3.2. Entity dictionary

In general, DBLP RDF contains named entities (in terms of URIs) of type publication and author. As described in Section 3, we collect descriptive information for the entities in order to disambiguate ambiguous surface forms.

For publications, we do not generate or collect alternative titles. Our dictionary contains the exact title as contained in the DBLP RDF KG. We assume that titles of publications are referenced in questions as exact and complete as possible. As the DBLP RDF KG does not contain references of publications to other publications, we did not include a popularity measure for publications in the dictionary. As context information, we collected the author names and the publisher for each publication.

In some cases, DBLP QuAD references authors in the questions only using their first name or even middle name. In other cases, one of the names is abbreviated and the others are not. Therefore, we generate an extensive set of alternative labels from their names. For instance, the author *Antonio Manuel Fernandez Villamor* has 21 entries in our dictionary with labels generated from the combination of parts of his name, as e.g. *antonio manuel fernandez v.*, *a. m. f. villamor*, or *antonio m. fernandez villamor*. Overall, our dictionary contains almost 30 million entries for DBLP authors. For the popularity measure, we counted the number of publications for each author. The context information for authors consists of the titles of all their publications transformed to a set of keywords and filtered for stop words. Also, we added all names of co-authors and all publishers the author have published in.

Table 1 shows the overview of information collected for both entity types.

Table 1
Descriptive information and origin for named entities in the dictionary

	Publications	Authors
alternative labels	/	each part of the name separately
indegree / popularity measure	/	number of publications
contextual information	author names & publisher	set of keywords from publications & co-author names & publishers

3.3. Handling of publishers

DBLP QuAD contains questions mentioning publications by specific publishers, as e.g. *Has J. Florens published in Computer Graphics in the last 5 years?*. One could assume that *Computer Graphics* is the surface form of a named entity represented by an URI in the DBLP RDF KG. But, this is not the case. The respective SPARQL query must use the following triple to express that something has been published in *Computer Graphics*:

```
?x <https://dblp.org/rdf/schema#publishedIn> 'Computer Graphics' .
```

In the AMR graph, the publisher *Computer Graphics* is represented in a name node as a named entity as shown in the graph below:

```
(p / publish-01
      :ARG0 (p2 / person
            :name (n / name
                  :op1 "J."
                  :op2 "Florens"))
      :ARG1 (p3 / publication
            :name n
            :op1 "Computer"
            :op2 "Graphics")
      :polarity (a / amr-unknown)
      :time (b / before
            :op1 (n2 / now)
            :duration (t / temporal-quantity
                  :quant 5
                  :unit (y / year))))
```

Our procedure to map a surface form to the underlying knowledge graph includes a similarity search in case an exact match is not successful. All identified named entities will be tried to be mapped in the KG. For *Computer Graphics* we would receive publications with the same name, as e.g. `https://dblp.org/rec/books/daglib/0067138` or `https://dblp.org/rec/journals/cg/HillM92`. Therefore, we added a check if the surface form is included in the labels of publishers. If so, the node is excluded from the mapping process [2].

4. Evaluation

As described in [1], we evaluated various combinations of parameters and settings for the disambiguation of named entities on the datasets QALD-9 and LC-QuAD 2.0. The best performing parameter settings for both datasets are the following:

- only the parent node (of the surface form) as descriptive context information
- weighting all disambiguation scores evenly (for QALD-9)
- weighting the popularity score and the label score higher than the context score (for LC-QuAD 2.0)

For DBLP-QuAD, we took part at the Scholarly QALD challenge 2023[3] and evaluated our approach on the test datasets of the development and the final phase. The settings as stated above for QALD-9 and LC-QuAD 2.0 did not achieve satisfying results on the DBLP-QuAD test datasets. With these settings, we achieved a maximum recall of 0.74 and maximum precision of 0.61. Therefore, we evaluated different weighting settings and achieved the best results on using only the context score – ignoring the popularity measure and the label score. In addition, we changed the context information from only the parent node to all name nodes in the AMR graph of the question and could increase recall and precision with this adjustment. Finally, we

[2]Publications stemming from the placeholder dictionary as described in Section **??** are flagged to be included in the mapping even though they share their name with a publisher.
[3]https://kgqa.github.io/scholarly-QALD-challenge/2023/

expanded the context information used for calculating the context score by using all words from the question filtered by stop words. Thereby, we achieved the best result overall with a recall of 0.926 and precision of 0.76 (F_1= 0.8353) on the test dataset of the final phase.

The following example depicts a typical disambiguation case. The question *Did Xiao L. publish in INTERSPEECH in the last 2 years?* from the DBLP-QuAD train dataset contains the highly ambiguous reference to an author *Xiao L.*. There are 93 authors with the abbreviated name *Xiao L.* listed in DBLP RDF. The author with most publications (as in our defined popularity measure) is a person with the name Xiao Liu[4] with 152 publications[5]. But the author required for the query as in the train dataset is a person with the name Xiao Li[6] with only 36 publications. In this case the context *INTERSPEECH* tips the scales and leads to the correct author. But, the dataset also contains many questions where authors are mentioned without any context. For all-purpose questions, such as *Did Heinrich Heine die in Paris?*[7], the most popular named entity with the name *Paris* would come into mind and is referenced as correct named entity in the respective QA dataset. But obviously, this rule does not apply for named entities in the DBLP-QuAD dataset and maybe also in general in datasets of very specific domains[8].

Therefore, the lessons learned from the entity linking challenge on DBLP-QuAD questions are the following:

- an extensive list of alternative labels for authors is essential
- the descriptive information of entities in the dictionary has to be collected thoroughly
- the context from the question is the most important information for the disambiguation

Hence, in this case for the DBLP-QuAD dataset, context matters even more than for datasets of a general domain.

5. Summary

We presented examination and adapted approach on entity linking for the scholarly domain using AMR graphs and an extensive entity dictionary. We showed that major adaptations are required for the scholarly domain and especially DBLP-QuAD compared to the original approach based on all-purpose knowledge graphs, such as Wikidata and DBpedia.

With this approach, we achieved a recall of over 92% and precision of over 76% and thereby winning the Scholarly QALD challenge 2023 on entity linking.

Our examination shows the importance of context for the disambiguation in a specific domain and especially when very ambiguous surface forms have to be resolved. We assume that this might due to the specific domain of the KG and the QA dataset. But, we need to further examine this issue on other domain-related KGs and datasets.

[4] https://dblp.org/pid/82/1364-4
[5] at least for our status of DBLP RDF
[6] https://dblp.org/pid/66/2069-6
[7] Question with uid 4303 in LC-QuAD 2.0 train dataset
[8] which needs to be proven on more datasets

References

[1] N. Steinmetz, Entity linking for KGQA using AMR graphs, in: C. Pesquita, E. Jiménez-Ruiz, J. P. McCusker, D. Faria, M. Dragoni, A. Dimou, R. Troncy, S. Hertling (Eds.), The Semantic Web - 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28 - June 1, 2023, Proceedings, volume 13870 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 122–138. URL: https://doi.org/10.1007/978-3-031-33455-9_8. doi:10.1007/978-3-031-33455-9_8.

[2] D. Milne, I. H. Witten, Learning to link with wikipedia, in: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 509–518. URL: https://doi.org/10.1145/1458082.1458150. doi:10.1145/1458082.1458150.

[3] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, Dbpedia spotlight: Shedding light on the web of documents, in: Proceedings of the 7th International Conference on Semantic Systems, I-Semantics '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 1–8. URL: https://doi.org/10.1145/2063518.2063519. doi:10.1145/2063518.2063519.

[4] P. Ferragina, U. Scaiella, Fast and accurate annotation of short texts with wikipedia pages, IEEE Software 29 (2012) 70–75. doi:10.1109/MS.2011.122.

[5] A. Sakor, K. Singh, A. Patel, M.-E. Vidal, Falcon 2.0: An entity and relation linking tool over wikidata, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 3141–3148. URL: https://doi.org/10.1145/3340531.3412777. doi:10.1145/3340531.3412777.

[6] K. Shivashankar, K. Benmaarouf, N. Steinmetz, From graph to graph: Amr to sparql, in: Proceedings of the 7th Natural Language Interfaces for the Web of Data (NLIWoD) co-located with the 19th European Semantic Web Conference (ESWC 2022), Hersonissos, Greece, May 29th, 2022, volume 3196 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: http://ceur-ws.org/Vol-3196.

[7] S. Neelam, U. Sharma, H. Karanam, S. Ikbal, P. Kapanipathi, I. Abdelaziz, N. Mihindukulasooriya, Y. Lee, S. K. Srivastava, C. Pendus, S. Dana, D. Garg, A. Fokoue, G. P. S. Bhargav, D. Khandelwal, S. Ravishankar, S. Gurajada, M. Chang, R. Uceda-Sosa, S. Roukos, A. G. Gray, G. Lima, R. Riegel, F. P. S. Luus, L. V. Subramaniam, SYGMA: system for generalizable modular question answering overknowledge bases, CoRR abs/2109.13430 (2021). URL: https://arxiv.org/abs/2109.13430. arXiv:2109.13430.

[8] P. Kapanipathi, I. Abdelaziz, S. Ravishankar, S. Roukos, A. G. Gray, R. F. Astudillo, M. Chang, C. Cornelio, S. Dana, A. Fokoue, D. Garg, A. Gliozzo, S. Gurajada, H. Karanam, N. Khan, D. Khandelwal, Y. Lee, Y. Li, F. P. S. Luus, N. Makondo, N. Mihindukulasooriya, T. Naseem, S. Neelam, L. Popa, R. G. Reddy, R. Riegel, G. Rossiello, U. Sharma, G. P. S. Bhargav, M. Yu, Leveraging abstract meaning representation for knowledge base question answering, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, Association for Computational Linguistics, 2021, pp. 3884–3894.

[9] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Zero-shot entity linking with dense entity retrieval, in: EMNLP, 2020.