Proposal for PORQUE, a Polylingual Hybrid Question Answering System*

Victor Mireles[1,*], Artem Revenko[1], Nikit Srivastava[2], Daniel Vollmers[2], Anna Breit[1] and Diego Moussallem[2]

[1] *Semantic Web Company GmbH*
[2] *Paderborn University*

Abstract

Organizations can benefit from integrating multilingual information from both textual and structured sources, and from its retrieval by means of Question Answering (QA) systems. Hybrid QA approaches, capable of finding answers in both documents and KGs, usually rely on translating textual sources into KG statements or vice-versa, and are often not leveraging the whole extent of a graph or the richness of the natural language text. Here we propose PORQUE, a hybrid QA system that utilizes multilingual language models, graph embeddings and modern decoder models to generate answers in many languages based on information contained in multilingual textual corpora and multilingual KGs. Of novelty is the hybrid representation of information which, guided by existing work in KG-augmented NLP, allows a more complete exploitation of both KG and documents.

Keywords

Question Answering, Knowledge Graph, Multilinguality

1. Introduction

Question Answering (QA) provides an easy and intuitive way to retrieve information in which the user can query using natural language questions and receive answers composed from several sources at once, without the need to engage with the organization of data sources. In many scenarios, the data from which the answers are composed is scattered across a collection of text documents and a structured data source, necessitating the development of a *Hybrid QA* system. In this paper, we propose the architecture of a future system for the case in which structured data takes the form of a Knowledge Graph (KG), be it in-house developed or part of the public LOD Cloud[1]. Furthermore, we are interested in the case in which the sources for answers is multilingual, and aim to support user interaction with the QA system in a variety of languages.

SEMANTICS 2022 EU: 18th International Conference on Semantic Systems, September 13-15, 2022, Vienna, Austria

[*] Funded by project PORQUE, of the Eureka Eurostars programme, Grant Number E114154

[*] Corresponding author.
victor.mireles@semantic-web.com (V. Mireles); artem.revenko@semantic-web.com (A. Revenko); nikit.srivastava@upb.de (N. Srivastava); daniel.vollmers@uni-paderborn.de (D. Vollmers); anna.breit@semantic-web.at (A. Breit); diego.moussallem@upb.de (D. Moussallem)
0000-0003-3264-3687 (V. Mireles); 0000-0001-6681-3328 (A. Revenko); 0000-0002-5324-4952 (D. Vollmers); 0000-0001-6553-4175 (A. Breit)

[1] https://lod-cloud.net/

Modern QA systems make use of Machine Learning (ML) methods that are trained on question-answer pairs and can afterwards infer answers for new questions. In this setting, hybrid approaches can integrate data sources (textual and structured) before or after the inference step. The former, known as *early fusion*, make use of a common representation of data, regardless of sources, on which the ML inference step is executed to produce an answer. The alternative, known as *late fusion*, make use of several data-source specific ML systems to produce answers, which are then combined using some heuristic. The *late fusion* approaches have recently been shown to be less performing [1], in part because they cannot exploit the information in one source to select or process information in another.

Current *early fusion* hybrid QA systems can be categorized in two, depending on the nature of the common representation on which the ML component operates.

2. KG2Text approaches to Hybrid QA

Several systems exist that verbalize the content of a KGs into text, and then apply QA methods that work on documents. For example, TeKGen [2] uses a Language Model (LM) to verbalize the entirety of Wikidata into over 15M sentences (known as the KELM corpus). The authors combine this corpus with the Wikipedia textual corpus, and use the state of the art retrieval LM know as REALM, to tackle two QA benchmarks.Another such system is UniK-QA [3] in which the authors use a Fusion-in-Decoder approach [4] with the T5 model to generate answers based on a similarly combined corpus.

The main advantage of the KG2Text approach is that it allows reusing the powerful machinery of pre-trained LMs, in the structured QA task, as formulated in [3]. A challenge in using KG2Text methods consists in identifying the best way to verbalize the structured knowledge. Another challenge consists in exploiting all available knowledge as it might be necessary to restrict the available search space, for example, “to a high-recall 2-hop neighborhood of the retrieved entities” in [3]. To the best of our knowledge, all KG2Text approaches are language-specific.

3. Text2KG approaches to Hybrid QA

To leverage the variety of existing systems that can answer questions over graphs, several systems convert textual documents to statements in a KG which can optionally then be linked to other, pre-existing KGs.

Some approaches generate only question-specific graphs by first pre-selecting a set of documents. For example, in the EGQA system [5], the authors use Wikipedia as a text source to extract documents using vector similarity based on TF-IDF representations. Afterwards, they extract triples from these documents to construct a raw graph using NER methods. Likewise, QUEST [6], generates a question-specific *pseudo-KG* using Open IE techniques from many question-relevant documents. It hence relies on Group Steiner Trees (GST), to identify nodes and consider them as answer candidates. Another example is GRAFT-Net [1], which first pre-selects sentences from documents using a Lucene index, and then executes entity-linking on the retrieved documents as well as on the original question.

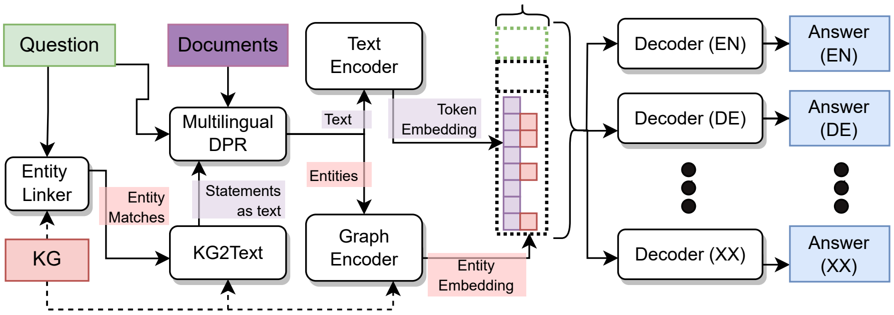

Figure 1: Our proposed architecture

Two systems overcome the limitations inherent in pre-selecting question-specific graphs, iteratively expanding them using queries to a larger graph. Uniqorn [7] a successor to QUEST, by using entity linking and PullNet [8], a successor of GRAFT-Net, by using a Graph Convolutional Neural Network (GCNN) to identify nodes that should be expanded.

Other approaches generate a large-scale, question independent, KG in a preprocessing step. One example is DELFT [9] which, in contrast to classical information extraction, builds a free-text knowledge graph from Wikipedia. DELFT's advantage comes from the high coverage of its KG which contains more than double that of DBpedia relations.

In general, Text2KG based method take similar approaches to generating answers: doing graph operations (e.g. querying or GSTs) to select a set of nodes that constitute the answer. While these approaches allow for answers coming from different sources, they neither provide answers in natural language nor utilize distributional semantics information contained in textual documents. The many advances that contextualized-word embeddings have brought to the QA domain are thus underutilized by Text2KG approaches.

4. PORQUE Approach

We propose a third approach, in which the ML inference step is executed over a shared, hybrid representation which does not make either of the two sources conform to the other. Our system, called PORQUE, combines KG embeddings with multilingual contextualized word embeddings allowing complete exploitation of the available knowledge sources.

PORQUE approaches question answering in an end-to-end manner, outputting a natural language answer which is not constrained to entities in the KG nor specific sentences in documents. The system is based on an encoder-decoder architecture (see Figure 1), and is partitioned into the modules described below, all of which are jointly refined on QA pairs. While a KG2Text module is present, it used only to pre-select sections of the KG which might be relevant, while the actual answer generation incorporates knowledge in the form of graph embeddings.

Entity Linker that takes in natural language text, extracts entities from it and links them to a multilingual KG. For the document corpus, the linking process is carried out offline, producing a lookup table. For the question, it is done online, outputting lists of tuples consisting of an entity URI and the token offset where it is located. It is based on tools like DBPedia Spotlight, Entity Fishing or PoolParty Semantic Suite, which are sensible to the input language.

KG2Text which converts triples in the KG into natural language sentences. This conversion is done to the list of entities produced by the Entity Linker from the input question, and the 2-hop surrounding graph. A table matching each of the generated sentences with the URIs of the entities involved is also kept. This verbalization, using methods such as those of [10] will be performed in English, since it is central in multilingual LMs.

Multilingual Dense Paragraph Retrieval (DPR), like DPR [11], performs a K-Nearest-Neighbor computation in the space of contextualized embeddings produced by a multilingual LM. It is trained on those paragraphs comprising the documents plus those generated from the KG. During inference, this module takes as input a question and produces: a list of *paragraphs* that are related to this question, and a list of entities mentioned in in them. The entities mentioned are recovered from the lookup tables from the KG2Text and Entity Linker modules.

Text Encoder takes as input a paragraph from the DPR module and produces, for each token, a vector representing a contextualized embedding. It is based on an existing multilingual text-embedding system (e.g. mT5 [12]).

Graph Encoder takes as input the URI of an entity and produces a vector representation. After experiments to determine their applicability to the QA task, an existing graph embedding system (e.g. ComplEx [13]) will be adopted.

Language-specific Decoder takes as input a *hybrid vector*, which is a combination of the output of both Encoder modules (discussed below), and outputs a natural language sentence. Depending upon on the desired language for an answer, the architecture can utilize the respective language specific instance of this module.

In PORQUE, answers are generated based on the *hybrid vectors* that represent information present both in the documents and in the graph. Each of these vectors corresponds to a token in a paragraph, and results from the combination (e.g., concatenation, see [14] for a discussion) of two components. The first is the contextualized, multilingual embedding of the token by the Text Encoder. The second is either i) the all-zeros vector in case the token is not part of any linked entity, ii) the graph-embedding as provided by Graph Encoder in case it is the start of an entity mention, or iii) the all-ones vector in case it is a subsequent token of an entity mention. This representation can also be generated for questions or documents containing no entities.

The Language-specific Decoders which are in charge of generating answers are presented with sequences of hybrid vectors. These sequences correspond to the token-sequences of each of the paragraphs (which can come either from the document corpus or from the verbalization of the KG) retrieved by the Multilingual DPR module, as well as the question itself.

By decoupling representation from any specific data type, one can leverage the multilingual capabilities already available in text and graph encoders. This reduces the need for Machine Translation systems, which have poor performance in domain-specific vocabularies [15], while producing answers in languages different to those of the underlying documents or KG.

Future Work The proposed system will be tested on industry use cases in the technical documentation and legal literature domains, as well as on standard benchmarks. Comparisons to other approaches, and analysis of the limitations of the method will then be published.

References

[1] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, W. W. Cohen, Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text, arXiv e-prints (2018) arXiv:1809.00782.

[2] O. Agarwal, H. Ge, S. Shakeri, R. Al-Rfou, Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training, in: Proceedings of the 2021 NAACL, ACL, Online, 2021, pp. 3554–3565.

[3] B. Oguz, X. Chen, V. Karpukhin, S. Peshterliev, D. Okhonko, M. Schlichtkrull, S. Gupta, Y. Mehdad, S. Yih, UniK-QA: Unified Representations of Structured and Unstructured Knowledge for Open-Domain Question Answering, arXiv e-prints (2020) arXiv:2012.14610.

[4] G. Izacard, E. Grave, Leveraging passage retrieval with generative models for open domain question answering, in: Proceedings of the 16th Conference of the European Chapter of the ACL: Main Volume, ACL, Online, 2021, pp. 874–880.

[5] G. Gu, B. Li, H. Gao, M. Wang, Learning to answer complex questions with evidence graph, in: APWeb-WAIM 2020, Proceedings, Part I, Springer-Verlag, 2020, p. 257–269.

[6] X. Lu, S. Pramanik, R. Saha Roy, A. Abujabal, Y. Wang, G. Weikum, Answering Complex Questions by Joining Multi-Document Evidence with Quasi Knowledge Graphs, arXiv e-prints (2019) arXiv:1908.00469.

[7] S. Pramanik, J. Alabi, R. Saha Roy, G. Weikum, UNIQORN: Unified Question Answering over RDF Knowledge Graphs and Natural Language Text, arXiv e-prints (2021) arXiv:2108.08614.

[8] H. Sun, T. Bedrax-Weiss, W. W. Cohen, PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text, arXiv e-prints (2019) arXiv:1904.09537.

[9] C. Zhao, C. Xiong, X. Qian, J. Boyd-Graber, Complex Factoid Question Answering with a Free-Text Knowledge Graph, arXiv e-prints (2021) arXiv:2103.12876.

[10] X. Li, A. Maskharashvili, S. Jory Stevens-Guille, M. White, Leveraging large pretrained models for WebNLG 2020, in: Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), ACL, 2020, pp. 117–124.

[11] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of EMNLP 2020, ACL, Online, 2020, pp. 6769–6781.

[12] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).

[13] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: ICML, PMLR, 2016, pp. 2071–2080.

[14] D. Moussallem, A.-C. Ngonga Ngomo, P. Buitelaar, M. Arcan, Utilizing knowledge graphs for neural machine translation augmentation, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 139–146.

[15] A. Perevalov, A.-C. N. Ngomo, A. Both, Enhancing the accessibility of knowledge graph question answering systems through multilingualization, in: 2022 IEEE 16th International Conference on Semantic Computing (ICSC), IEEE, 2022, pp. 251–256.