

A Universal Question-Answering Platform for Knowledge Graphs

Reham Omar*, Ishika Dhall*, Panos Kalnis[§], Essam Mansour*

*Concordia University, Canada
{fname}. {lname}@concordia.ca

[§]KAUST, Saudi Arabia
panos.kalnis@kaust.edu.sa

ABSTRACT

Knowledge from diverse application domains is organized as knowledge graphs (KGs) that are stored in RDF engines accessible in the web via SPARQL endpoints. Expressing a well-formed SPARQL query requires information about the graph structure and the exact URIs of its components, which is impractical for the average user. Question answering (QA) systems assist by translating natural language questions to SPARQL. Existing QA systems are typically based on application-specific human-curated rules, or require prior information, expensive pre-processing and model adaptation for each targeted KG. Therefore, they are hard to generalize to a broad set of applications and KGs. In this paper, we propose KGQAN, a *universal* QA system that does not need to be tailored to each target KG. Instead of curated rules, KGQAN introduces a novel formalization of question understanding as a text generation problem to convert a question into an intermediate abstract representation via a neural sequence-to-sequence model. We also develop a just-in-time linker that maps at query time the abstract representation to a SPARQL query for a specific KG, using only the publicly accessible APIs and the existing indices of the RDF store, without requiring any pre-processing. Our experiments with several real KGs demonstrate that KGQAN is easily deployed and outperforms by a large margin the state-of-the-art in terms of quality of answers and processing time, especially for arbitrary KGs, unseen during the training.

KEYWORDS

Natural Language Question Answering, Knowledge Graphs, RDF, Seq2Seq Models, Just-In-Time Entity and Relation linking

ACM Reference Format:

Reham Omar*, Ishika Dhall*, Panos Kalnis[§], Essam Mansour*. 2023. A Universal Question-Answering Platform for Knowledge Graphs. In *Proceedings of the 2023 International Conference on Management of Data (SIGMOD '23)*. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3588911>

1 INTRODUCTION

There is a steep rise in the availability of knowledge graphs (KGs) across various application domains. Examples include KGs about scientific publications, such as the Microsoft Academic Graph¹,

¹<https://makg.org/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

SIGMOD '23, June 18–23, 2023, Seattle, WA

© 2023 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/3588911>

```
1 PREFIX dbv:<http://dbpedia.org/resource/>
2 SELECT ?sea WHERE {
3   ?sea <http://dbpedia.org/property/outflow>
4     dbv:Danish_straits .
5   ?sea <http://dbpedia.org/ontology/nearestCity>
6     dbv:Kaliningrad . }
```

Figure 1: SPARQL query for question q^E . “city on shore” maps to `<http://dbpedia.org/ontology/nearestCity>`.

DBLP² and OpenCitations³; general-fact KGs, such as DBpedia⁴, Wikidata⁵ and YAGO⁶; knowledge related to bio-sciences (e.g., Bio2RDF⁷); or politics (e.g., the UK Parliament⁸ KG). Due to the simplicity of the RDF model, the powerful SPARQL query language and the extensions for inferencing and reasoning [5, 24], many KGs are stored in RDF data management systems [1, 2] that can be accessed remotely in the web as SPARQL endpoints.

Consider q^E = “Name the sea into which Danish Straits flows and has Kaliningrad as one of the city on the shore”, a question that appears in the popular LC-QuAD 1.0 [52] benchmark. Intuitively, q^E corresponds to a conjunctive SPARQL query consisting of two triples: (`?sea, flows, DanishStraits`) and (`?sea, cityOnShore, Kaliningrad`). In practice, assuming the endpoint is DBpedia, the query should be expressed as shown in Figure 1. Notice that the correspondence between elements of q^E (e.g., “city on shore”) and the actual SPARQL components (i.e., `<http://dbpedia.org/ontology/nearestCity>`) is not trivial.

It is impractical for the average user to know the schemata of the KGs, or the exact URIs of the entities and predicates to express the appropriate SPARQL queries. Hence, recent research [27, 28, 31, 64] investigated the automatic translation of natural language questions to SPARQL; the problem is known as question answering (QA). Existing QA systems utilize question understanding approaches based on semantic parsing and curated rules that are tailored to a specific KG; consequently, they are hard to generalize to broader application domains. Moreover, existing QA systems rely on prior knowledge for each target KG to link question phrases into KG vertices and predicates. For example, gAnswer [27, 64] and EDGQA [28] perform expensive pre-processing to generate indices for each KG, whereas NSQA [31] trains a specialized deep neural network on the target KG. This has several drawbacks: (i) the computational cost of pre-processing is substantial, especially for large KGs; (ii) QA is only available at those SPARQL endpoints whose owners choose

²<https://dblp.org/>

³<https://opencitations.net/sparql>

⁴<https://dbpedia.org/sparql>

⁵<https://query.wikidata.org/>

⁶<https://yago-knowledge.org/sparql>

⁷<https://bio2rdf.org/sparql>

⁸<https://api.parliament.uk/sparql>

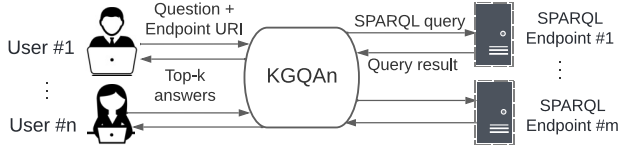


Figure 2: Users interact on-demand with KGQAN to submit a natural language question against an arbitrary KG. Question understanding is independent of the KG. The construction of the actual SPARQL query is performed in a JIT manner, through the publicly accessible API of the target KG.

to install and continuously maintain the QA system; and (iii) it is hard to generalize to arbitrary KGs.

In contrast to existing systems, we are inspired by web search engines that resolve user questions independently of any particular web site. We propose a novel approach, called KGQAN⁹, that forgoes the need of tailoring the QA system to each individual KG. KGQAN resides between the users and any available KG, as shown in Figure 2. KGQAN introduces novel approaches for *question understanding*, *linking*, and *filtering* of the final answers. We formalize question understanding as a text generation task that extracts abstract triple patterns from the natural language question. Prior to deployment, we train KGQAN on a diverse set of questions using a sequence-to-sequence (Seq2Seq) [50] deep neural network. The resulting model takes a question as input and extracts a set of triple patterns that represent the relations between entities, which are either explicitly mentioned, or are unknowns. With those triple patterns we construct the so-called phrase graph pattern (PGP), which is a formal abstract representation of the system’s understanding of the question, independently of any KG.

We also propose a just-in-time (JIT) linker that negotiates with the target KG the mapping between abstract elements of the PGP and the appropriate URIs to instantiate a well-formed SPARQL query. KGQAN submits requests via the publicly accessible SPARQL endpoint API and utilizes common built-in indices [44] that exist in all popular RDF stores (e.g., Virtuoso, Stardog, and Apache Jena). Linking is performed transparently via a semantic affinity model trained on general English text, without the need of any particular configuration or action by the owner of the KG. The SPARQL query is submitted to the KG and the answers are collected and filtered by KGQAN before being returned to the user. Filtering is based on constraints such as the expected domain, or data type of the answer; the constraints are identified during the construction of the PGP. Unlike existing QA systems, our filtering method does not need any prior knowledge about the KG.

We evaluate KGQAN on a variety of real KGs. In addition to DBpedia, which is used by the two popular QA benchmarks, LC-QuAD 1.0 [52] and QALD-9 [54], we also experiment with YAGO, the Microsoft Academic Graph and DBLP. In terms of answer quality (i.e., Macro F1 score) and question processing time, our experiments confirm that KGQAN outperforms existing QA systems by a large margin when queries are executed on arbitrary KGs, unseen during training. Interestingly, despite the fact that KGQAN is general and does not target any particular KG, it manages to perform almost

identically, or in some cases even outperform, the current state-of-the-art (SOTA) on the LC-QuAD 1.0 and QALD-9 benchmarks; note that the SOTA is optimized specifically for those benchmarks.

In summary, our contributions are:

- KGQAN, the first universal QA system that can process on-demand questions against arbitrary, previously unseen KGs.
- A novel formalization of question understanding as a text generation problem. We train Seq2Seq neural networks that learn how to extract a formal abstract representation from the natural language question, independently of the application domain and without any curated rules.
- A novel just-in-time linking and filtering approach that utilizes a generic semantic affinity model and the public API of the SPARQL endpoint. Our approach does not need prior knowledge of the KG and aims at maximizing the recall of the SPARQL query, before improving precision via filtering.
- A comprehensive evaluation using four real KGs from different domains. KGQAN performs similarly to the SOTA for the known benchmarks, but outperforms the SOTA by a large margin for previously unseen KGs.

The paper is organized as follows: Section 2 clarifies the necessary background and limitations of existing systems. Sections 3, 4, 5 and 6 present the architecture of KGQAN, as well as our query understanding, linking and filtering approaches. Section 7 contains the experimental evaluation. Section 8 discusses the related work and Section 9 concludes the paper.

2 BACKGROUND AND LIMITATIONS

QA systems commonly split the question answering process into three steps: (i) *question understanding*: extracts the entities and relations from the question and generates an abstract representation; (ii) *linking*: maps the abstract representation to the corresponding vertices and predicates in the targeted KG to construct a well-formed SPARQL query; and (iii) *filtering*: selects the most relevant answers, based on some constraints. In this section we discuss three existing QA systems, gAnswer [27, 64], EDGQA [28] and NSQA [31], summarized in Table 1. These systems highlight the most successful approaches, as they represent the SOTA in terms of accuracy (i.e., F1 score) on the LC-QuAD 1.0 and QALD-9 benchmarks.

2.1 Question Understanding

The question understanding step identifies mentioned entities, unknowns and relations in the natural language question. It also generates an abstract formal representation (typically, a graph) of their inter-dependencies. Figure 3 shows the expected result for our running example q^E .

gAnswer uses the Stanford dependency parser [12] to map a question into a syntactic dependency tree. The tree is traversed to generate an abstract graph representation of the query, called semantic query graph. This is achieved by a set of static heuristic linguistic rules that are curated on the set of the QALD-9 training questions; for instance, question words starting with “wh*” (e.g., “what”) are assumed to represent unknowns. gAnswer also depends on a predefined set of synonyms [41]. For our running example, gAnswer generates a lot of unnecessary and erroneous triples, such as (Kaliningrad, city of, city).

⁹KGQAN: Knowledge Graph Question ANswering platform.

Table 1: A Comparative Analysis of SPARQL-based QA systems achieved highest F1 scores in QALD-9 and/or LC-QuAD 1.0

	Question Understanding	Linking	Filtering
gAnswer [27, 64]	Dependency parsing	Index look up	n/a
EDGQA [28]	Constituency parsing	Dexter, EARL, Falcon	By index type
NSQA [31]	AMR parsing, BERT-based	BLINK, SEM-Rel	n/a
KGQAN (ours)	Seq2Seq text generation	RDF engine + Semantic similarity	By predicted answer type

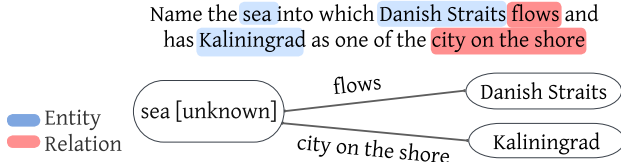


Figure 3: The understanding of question q^E identifies Danish Straits and Kaliningrad as mentioned entities; sea as an unknown, flows and city on the shore as relations, resulting in a graph with two triple patterns.

EDGQA utilizes the Stanford core NLP parser [10] to construct the constituency parsing tree of the question. Then, it generates a rooted acyclic graph, called entity description graph, by iteratively decomposing the parsing tree. The process is based on rules that are tailored to the LC-QuAD 1.0 and QALD-9 benchmarks. For our running example, EDGQA erroneously detected [20] Danish Straits as a relation, flow as an entity, and generated this triple pattern: (unknown, Danish Straits, flow).

NSQA parses the question into a directed rooted acyclic graph, called abstract meaning representation. Parsing is based on a deep neural symbolic framework. The neural network is trained on verb-oriented relations gathered from Propbank. The training set also consists of a set of curated trees for different questions. The accuracy of the model depends on the diversity of these trees. It is unclear¹⁰ whether the model can detect relations based only on noun phrases, such as “city on shore”. The resulting graph has different granularity than SPARQL; thus, the model has to be adjusted [31] to work with new domains.

2.2 Entity and Relation Linking

In QA systems, linking is a challenging task [33] that attempts to map entities and relations extracted from a question to the corresponding vertices and predicates in a KG. For our running example q^E on DBpedia, the expected output is:

Phrase	URI
Danish Straits	<http://dbpedia.org/resource/Danish_straits>
Kaliningrad	<http://dbpedia.org/resource/Kaliningrad>
flows	<http://dbpedia.org/property/outflow>
city on the shore	<http://dbpedia.org/ontology/nearestCity>

Existing systems formalize linking as: (i) a lookup in an inverted index that maps synonyms to relevant vertices and predicates of the target KG; or (ii) a keyword search on a database of tagged

¹⁰NSQA is a proprietary IBM project, not available to us to reproduce the results, or test our running example q^E ; our discussion is based on information from [31].

documents constructed from the vertices and predicates of the target KG; or (iii) a deep learning-based transformation, defined on the vector space of the target KG. These methods are KG-specific and require expensive pre-processing for each target KG.

gAnswer links entities via the crossWikis dictionary [49], constructed from a web crawl. Each entity in the dictionary is mapped to a corresponding vertex in the target KG with a confidence score. For relation linking, gAnswer constructs a dictionary that maps relation mentions to their corresponding predicates in the target KG. For example, relation “starred by” is mapped to predicate starring in a particular KG. The mapping is generated by general NLP [35] methods, or by using n-grams [21].

EDGQA utilizes three indexing systems, namely Falcon [47], EARL [18] and Dexter [6], to index vertices and predicates of the target KG. EDGQA ensembles the results from the three systems to link entities to KG vertices; then uses those to limit the search space for relation linking [40]; and finally utilizes a BERT-based [13] model to rank the list of predicates. The aforementioned indexing systems process a target KG by extracting vertices and predicates with their associated descriptions. For example, Falcon uses standard predicates #label and #altLabel to collect descriptions and synonyms. On those, Falcon performs part-of-speech tagging and n-gram extraction to generate tagged documents, in order to enable keyword search during linking. If a KG does not contain #label predicates, an appropriate indexing predicate must be selected manually for each vertex type, e.g., authorName can be chosen to index vertices of type author.

NSQA is slightly different, since it identifies entities and relations during linking. First, it employs Blink [61] to generate an embedding of the target KG vertices in a vector space. Then, it receives the abstract graph from the previous phase and detects entities via a BERT-based neural model, trained on 3,651 questions from LC-QuAD 1.0 with manually annotated mentions. The detected entities are embedded in the same space as the KG and are linked to their nearest-neighbor vertices. For relation linking, NSQA uses SemRel [36], a transformer-based neural model, which is trained to predict the relation. The model returns a list of candidate relations from the set of relations in the KG and ranks them. In NSQA, building the vector space, that is, generating the embedding for the target KG vertices and predicates, dominates the cost of the pre-processing phase.

2.3 Filtering

Some QA systems, such as EDGQA, include a step for filtering the potential answers by the RDF engine. For the filtering step, existing systems build prior knowledge during a pre-processing phase of all node types/classes in a KG. The filtering is done by linking unknown (e.g., ?sea) to a vertex type in the target KG, such as (?sea, a, dbo:Sea), if such a type exists. The SPARQL query

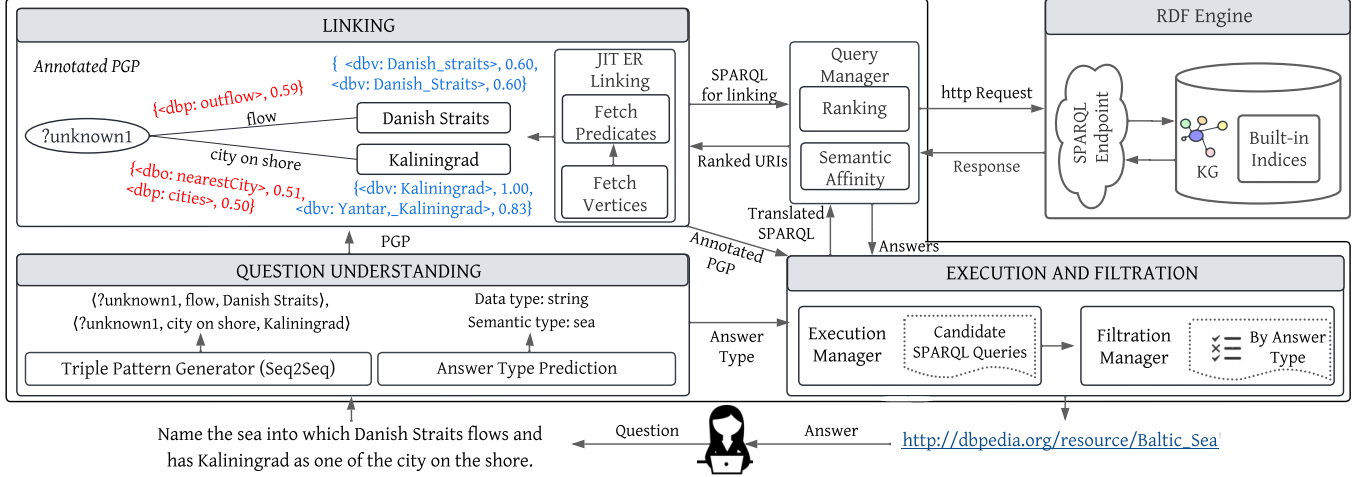


Figure 4: KGQAN’s architecture includes three phases: a) Question Understanding: a PGP is constructed using our Triple Pattern Generator (seq2seq model), b) Linking PGP to KG: JIT ER linking module annotates the PGP with relevant vertices and predicates from a KG, and c) Execution and Filtration: creating and executing SPARQL queries, and post-filtering their results.

amended with the type constraint is sent for execution; therefore, the received answers are pre-filtered.

3 KGQAN ARCHITECTURE

Similar to existing approaches, KGQAN implements three distinct phases: question understanding, linking, and execution with filtration, as shown in Figure 4. However, in contrast to existing approaches that require prior knowledge and expensive pre-processing for each target KG, KGQAN is a universal solution that works with arbitrary KGs from a variety of application domains.

Question understanding. KGQAN formalizes question understanding as a text generation task. We employ a pre-trained Seq2Seq language model, such as BART [32] or GPT-3 [4]. The original model understands general English language text; we amend its training with a set of typical questions and their corresponding translation to abstract triple patterns. Training is performed independently of any KG. Consequently, the model accepts general English questions and generates a sequence of abstract triple patterns. Figure 4 shows that, for our running example q^E , the seq2seq model generates two triples: $\langle ?unknown1, flow, Danish\ Straits \rangle$ and $\langle ?unknown1, city\ on\ shore, Kaliningrad \rangle$. The set of all triples correspond to an abstract graph representation, called phrase graph pattern (PGP).

KGQAN also predicts the expected data and semantic type of the answer. We train a three-layer neural network; its input is the question in English and the output is the expected data type of the answer, that is, date, numerical, boolean, and string. In the latter case, we also predict the semantic type, based on the context. For our example question, the predicted data type is string, whereas the predicted semantic type is “sea”.

Linking. Intuitively, linking corresponds to semantic match between elements of the abstract PGP graph and the actual data in the target KG. KGQAN does not have any prior information about the target KG; therefore, the match must be performed in a just-in-time (JIT) manner, during question answering. This allows KGQAN to support arbitrary KGs. Since most RDF engines do not support semantic search, we split this process between KGQAN and the

RDF engine. KGQAN submits, through the standard SPARQL API, a set of text containment queries for the phrase associated with each entity in the PGP; these are answered by the built-in indices [44] of the RDF engine. The answers, which are URIs of potential semantic matches to the PGP phrases, are assigned by KGQAN a semantic affinity score, based on an existing generic word embedding model. The top- k matches are then used to query again the target KG (by a standard SPARQL query), this time in order to acquire potential semantic matches for the predicates of the PGP. The result of this process is an *annotated* PGP. For our running example in Figure 4, entity Kaliningrad has two matches, $dbv:Kaliningrad$ with score 1.00 and $dbv:Yantar_Kaliningrad$ with score 0.83; whereas relation $flow$ has one match, $dbp:outflow$ with score 0.59.

Execution and filtration. KGQAN traverses the annotated PGP to generate all combinations of well-formed candidate SPARQL queries that can be semantically equivalent to the user question, for the target KG. The top- k most promising queries, based on a semantic affinity score, are sent to the RDF engine. The answers are collected by KGQAN and are filtered based on the aforementioned expected data and semantic type. Note that, while the execution of multiple candidate queries increases recall, the filtering step crucially improves precision. In contrast to existing systems that embed a KG-specific filter within the SPARQL query, our filtering method is applied at a post-processing phase, independently of the KG; therefore, our method is applicable to arbitrary KGs. For our running example, the user receives http://dbpedia.org/resource/Baltic_Sea as final answer.

4 KGQAN QUESTION UNDERSTANDING

This section introduces our novel formalization of the question understanding problem as a text generation task. Given a question q in plain English, we generate a sequence of triple patterns $\langle entity_i^a, relation_i, entity_i^b \rangle$ that represent our formal understanding of q ; this is achieved by a Seq2Seq deep neural network. All components of the triples are either phrases from q , or unknowns (i.e., variables); therefore question understanding is independent of any specific

domain or KG. Unknowns, in particular, are first-class citizens; we detect their semantic relations with mentioned entities, and predict their expected data and semantic type. Note that, our work is related to relation triple extraction [9, 58] that refers to the construction of a KG from long text; however, the existence of unknowns renders our task more challenging.

4.1 Phrase Triple Patterns Extraction

Question q in plain English consists of a sequence of words $(q_1, q_2, q_3, \dots)$, where q_i is the i th word. Our task is to generate a sequence of triples that capture the semantics of q . Formally:

Definition 4.1 (Phrase Triple Pattern Extraction). Let $q = (q_1, q_2, q_3, \dots)$ be a question. Generate a sequence of triples $TP(q) = (tp_1, tp_2, \dots)$. Each $tp_i \in TP(q)$ has the form $tp_i = \langle e_i^a, r_i, e_i^b \rangle$, where e_i^a and e_i^b are either entity phrases in q , or unknowns; and r_i is a relation phrase in q .

4.1.1 Triple extraction by a Seq2Seq transformer. We model our text generation task using a Seq2Seq pre-trained language model (PLM). The input query q is in plain text. For compatibility with the seq2seq model, the output $TP(q)$ should also be represented as text [50]. Let $y = F_{txt}(TP(q))$ be the text representation. F_{txt} transforms each $tp_i \in TP(q)$ into plain text by annotating e_i^a, e_i^b, r_i as “EntityA”, “EntityB” and “Relation”, respectively. Our example query q^E results in two triples: $TP(q^E) = (\langle ?unknown1, flow, Danish\ Straits \rangle, \langle ?unknown1, city\ on\ shore, Kaliningrad \rangle)$, that correspond to the following triple patterns:

```
[Relation(label="flow"),
  EntityA(label="Name", category=variable, varID=1),
  EntityB(label="Danish Straits", category=entity)],
[Relation(label="city on shore"),
  EntityA(label="Name", category=variable, varID=1),
  EntityB(label="Kaliningrad", category=entity)]
```

We consider two categories of PLMs: (i) encoder-decoder, such as BART [32]; and (ii) decoder-only, such as GPT-3 [4]. These PLMs support dynamic-length sequence generation [53]. Figure 5 illustrates our training process using an encoder-decoder PLM. The encoder receives q as input and applies a self-attention mechanism to generate a vector representation (embedding) of q . The encoder-embeddings $F^e(q)$ encapsulate various features of q that emphasize its most important tokens. The decoder learns to generate triple patterns as a sequence of tokens, where predicting token i is based on $F^e(q)$ and all previous tokens. The decoder uses two attention mechanisms: (i) self-attention on its input y , and (ii) cross-attention between $F^e(q)$, and y [56]. If a decoder-only PLM is used, q is directly provided as input to the decoder, which is trained to generate y using only a self-attention mechanism on q [4].

4.1.2 Training dataset for triple pattern extraction. The Seq2Seq model, either BART or GPT-3, is pre-trained with general English text. To be used for question understanding, we must perform additional training for triple extraction, as shown in Figure 5. To achieve this, we prepare a manually annotated training dataset using 1,752 questions collected from the training datasets of LC-QuAD 1.0 [52] and QALD-9 [54]. Each question q is annotated with

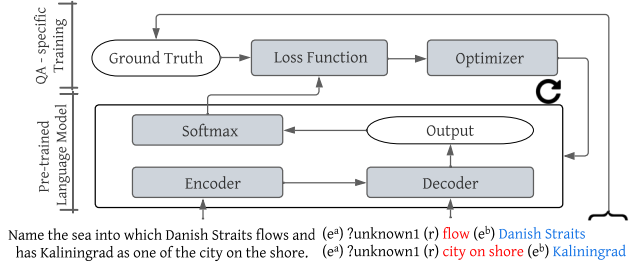


Figure 5: Training our Seq2Seq model with a question and a sequence of triple patterns as inputs to the encoder and decoder, respectively. Training is based on 1,752 manually annotated questions.

an appropriate set of triples $TP(q)$. For each q , the main steps of the annotation process are:

- Identify the number of phrase triple patterns $|TP(q)|$.
- Detect the triple patterns that share the same unknown.
- Assign a unique identifier for each unknown.
- For each $\langle e_i^a, r_i, e_i^b \rangle \in TP(q)$, extract from q the entity and relation phrases that correspond to e_i^a, e_i^b and r_i , respectively.

We design our training dataset to include questions with different numbers of unknowns and triple patterns. In our annotation, we assume one main unknown¹¹ (i.e., intention) for which we need to provide an answer. Additional unknowns may exist, but are treated as intermediate variables. Distinguishing between main and intermediate unknowns helps KGQA at the execution and filtration phase. For example, intermediate variables can be used in the WHERE clause, but not in the SELECT clause of the generated SPARQL query.

We consider various cases for the syntactic and semantic expression of entities and relations. For example, an entity may be a named entity, such as “Danish Straits”; or an entity mention, such as “capital region”. A relation may be a verb, such as “flows”; a verb with adverb, such as “work out”; or a noun phrase, such as “city on shore”. Our dataset allows the Seq2Seq model to learn how to distinguish between entity and relation phrases, without any additional semantic parsing, or part-of-speech tagging.

Our dataset can be amended in the future with additional representative cases. However, it is already general enough to allow the Seq2Seq model to extract triples from a variety of domains. For example, although our training involves general facts from DBpedia, without any questions about publications, it can understand questions related to DBLP, as shown in our experimental evaluation. Also note that the annotated dataset does not contain the actual SPARQL query, since this would depend on the target KG, which is not known during training.

4.2 Phrase Graph Pattern (PGP)

Let q be a question and $TP(q)$ be the corresponding set of triple patterns. We define the phrase graph pattern for q as:

¹¹Both QALD-9 and LC-QuAD 1.0 benchmarks do not include questions with two intentions. In future work, we plan to extend KGQA to support questions with two intentions, e.g., when and where did Covid-19 start?

Definition 4.2 (Phrase Graph Pattern (PGP)). Let $\mathcal{E}, \mathcal{R}$ be a set of nodes and edges, respectively. The phrase graph pattern of q is an undirected graph $PGP(q) = (\mathcal{E}, \mathcal{R})$, such that: for every triple $\langle \text{entity}_i^a, \text{relation}_i, \text{entity}_i^b \rangle \in TP(q)$ there is a node $e_i^a \in \mathcal{E}$ with label entity_i^a ; a node $e_i^b \in \mathcal{E}$ with label entity_i^b ; and an edge $(e_i^a, e_i^b) \in \mathcal{R}$ with label relation_i .

Intuitively, $PGP(q)$ connects the set of triple patterns in $TP(q)$ into a graph that represents the formal understanding of question q , as shown in Figure 4. Note that, although RDF data form directed graphs, $PGP(q)$ is undirected. The reason is that the PGP is not aware of the target KG. Therefore, at this point, we do not know yet the predicates that appear in the target KG, nor their direction. Depending on the question, the constructed PGPs may form different shapes, such as star, or path query graphs. Following the different categories of questions in [17], the current implementation of KGQAN can support a wide range of questions, including: single fact, single fact with type, multi-fact, and Boolean questions.

4.3 Data and Semantic Type Prediction

As explained above, the resulting PGP contains one main unknown. KGQAN predicts the expected data type and the semantic type of the unknown; this information will be used after the query execution, in order to improve the precision of the answers. KGQAN uses only the user question q to predict the data and semantic types, irrespectively of the target KG.

The expected data type can be date, numerical, boolean, or string. We define data type prediction as a classification task. We train a deep neural network using the training dataset in the QALD-9 benchmark, since its questions are already annotated with the data type. If the expected data type is string, then we also predict the semantic type (e.g., person, book, etc). We use the following heuristic: the first noun in the question is the semantic type. We utilize the AllenNLP constituency parser [26] to acquire the part-of-speech tags in the question, and extract the first phrase whose tag is noun. This heuristic has impact only on the accuracy of KGQAN’s post-filtering step; question understanding and linking are not affected. For our running example q^E , KGQAN predicts the data and semantic types to be string and “sea”, respectively.

5 KGQAN LINKING

For a question q , linking receives $PGP(q)$ from the question understanding phase, and generates an annotated version $AGP(q)$ of the graph. Vertices and edges in $AGP(q)$ are annotated with the URIs from the target KG that are semantically closest to the meaning of q . Since the target KG is not known in advance, KGQAN implements linking at query time, in a just-in-time manner using a set of light-weight SPARQL requests. This allows KGQAN to work with arbitrary KGs, without any pre-processing.

5.1 Entity Linking

Let l_n be the label of a node n in $PGP(q)$. Let d_v be the description (i.e., a literal of type string) of a vertex v in the target knowledge graph KG . Let $S(l_n, d_v)$ be the semantic affinity score between l_n and d_v ; refer to Section 5.4 for the definition of $S(\circ, \circ)$. Intuitively, for each node n we need to go through all vertices v in KG ; compute

their semantic affinity to n ; and select the top- k matches. These are the *relevant* vertices that are added as annotation on n in the annotated graph $AGP(q)$.

There are two practical issues: First, how to extract the description d_v of node v in KG . Some KGs have human-readable URIs and clearly marked descriptions via the `rdf:label` predicate. For example, the entry for Princess Diana in DBpedia contains triple $\langle \text{http://dbpedia.org/resource/Diana_Princess_of_Wales}, \text{rdf:label}, \text{“Diana, Princess of Wales”} \rangle$. We can check with a SPARQL ASK query if KG contains `rdf:label` predicates, in which case we can directly retrieve the descriptions. Other KGs, however, are more cryptic. For example, the entry for Jim Gray in the Microsoft academic graph, is `mag:2279569217`. Thankfully, there is a triple $\langle \text{https://makg.org/entity/2279569217}, \text{foaf:name}, \text{“Jim Gray”} \rangle$ from which we can retrieve the description. However, `foaf:name` is an arbitrary predicate that may differ for other entities. Therefore, in general we retrieve triples that connect KG vertices to literals of type string via any predicate.

The second issue is that typical RDF engines do not support semantic search. Consequently, function $S(\circ, \circ)$ must be computed remotely, at the KGQAN site. Obviously, we do not want to transfer a large amount of (possibly irrelevant) data from KG to KGQAN. We use the following heuristic: label l_n consists of a sequence of words (e.g., “Danish Straits”). We request from KG only those vertices v that are connected with a literal d_v via any predicate p , such that d_v contains any combination of words from l_n . The heuristic translates to the following SPARQL query, where $Q(l_n)$ represents a disjunctive boolean expression for all words in l_n :

```
QUERY: potentialRelevantVertices(l_n, maxVR)
1 SELECT DISTINCT ?v ?d_v
2 WHERE {
3   ?v ?p ?d_v .
4   ?d_v <bif:contains> Q(l_n) . }
5 LIMIT maxVR
```

All modern RDF engines, such as Virtuoso, Stardog, and Apache Jena, construct by default full-text indices to enable text search [44]; therefore, `potentialRelevantVertices` can be executed efficiently. Since the query may still return numerous matches, we heuristically limit the result size to `maxVR`; in our experiments, we set `maxVR` = 400. The query assumes Virtuoso as the RDF engine. Other engines may expose a slightly different API; for example, for Stardog we replace `<bif:contains>` with `<stardog:textMatch>`. We formally define the set of relevant vertices as:

Definition 5.1 (Relevant vertices $\mathcal{R}_V$). Given a question q and a target knowledge graph $KG = (V, P)$ considered as a set of triples. Let n be a node in $PGP(q)$ with label l_n . Let $T_d = \{ \langle v, p, d_v \rangle : v \in V, p \in P \}$ be a subset of triples from KG , such that description d_v is a literal of type string, and l_n is fully or partially contained in d_v . Let $T_v = \{ \langle v, S(l_n, d_v) \rangle : \langle v, p, d_v \rangle \in T_d \}$, where $S(\circ, \circ)$ represents the semantic affinity score. The set of relevant vertices $\mathcal{R}_V(n, KG)$ for node n in KG , is the subset of T_v that contains all pairs with the top- k affinity scores.

Observe that the `potentialRelevantVertices` query, which is executed at the target KG, is a heuristic that retrieves fast a

Algorithm 1 KGQANEntityLink

Input: n : a node in $PGP(q)$, KG : target knowledge graph, $maxVR$: max fetched vertices, k : number of vertices

Output: n annotated with relevant k vertices $\mathcal{R}_V(n, KG)$

```
1: if  $n.type$  is “unknown” then                                 $\triangleright n$  is a variable
2:   return  $n.\mathcal{R}_V \leftarrow \emptyset$ 
3: end if
4:  $T_d \leftarrow \text{potentialRelevantVertices}(n.l_n, maxVR)$   $\triangleright$  SPARQL to
    $KG$ 
5:  $T_v \leftarrow \emptyset$ 
6: for every  $\langle v, d_v \rangle \in T_d$  do
7:    $T_v \leftarrow T_v \cup \langle v, S(n.l_n, d_v) \rangle$   $\triangleright$  compute semantic affinity
8: end for
9: return  $n.\mathcal{R}_V \leftarrow$  all pairs from  $T_v$  with top- $k$  affinity score
```

relatively large and potentially inaccurate set of vertices. For PGP node Kaliningrad in our running example q_E , it returns both dbv:Kaliningrad and dbv:Yantar, _Kaliningrad (see Figure 4). The semantic affinity score, which is computed at the KGQAN site, identifies dbv:Kaliningrad as the top match.

Algorithm 1 summarizes the entity linking process for a single node n ; the algorithm must be called for every node $n \in PGP(q)$. In line 1 we check if n is an unknown (i.e., variable), in which case there will be no relevant vertices at this phase (line 2). Line 4 executes the potentialRelevantVertices SPARQL query at the target RDF engine. Lines 5-8 compute for each returned vertex its semantic affinity to n , and store the resulting $\langle vertex, score \rangle$ pairs in T_v , as explained in Definition 5.1. Finally, line 9 selects from T_v the pairs with the top- k score to construct the set of relevant vertices $\mathcal{R}_V(n, KG)$.

5.1.1 Complexity. Algorithm 1 is executed $|\mathcal{E}|$ times, where $\mathcal{E}$ is the set of nodes in $PGP(q)$; see Definition 5.3. The cost is dominated by the SPARQL query at line 4. Let c_{rdf} be the cost of performing keyword search in an RDF engine. c_{rdf} is determined by the indices and methods implemented in the particular engine. For example, Apache Jena implements full text search by either Lucene [30], or Elastic search. The for-loop in line 6 is executed a constant number of times, because it is constrained by parameter $maxVR$ in the SPARQL query. The cost of the semantic affinity calculation in line 7 depends on $|l_n| \cdot |d_v|$. However, in practice both of these sets contain only a few words, so the cost can be considered constant. The top- k computation in line 9 is also constrained by $maxVR$, so it is constant. The resulting complexity of Algorithm 1 is $O(c_{rdf}|\mathcal{E}|)$.

5.2 Relation Linking

Let p be a predicate in the target knowledge graph KG . Since KG is directed, p must be connected to some vertex $v \in KG$ either as an outgoing, or incoming edge. If we consider KG in its triple representation, there exists triple $\langle v, p, ?obj \rangle$, or $\langle ?sub, p, v \rangle$, respectively. Let d_p be the description (i.e., a human-readable string) of p . In some KGs, the URIs of the predicates contain human-readable text, such as dbp:spouse in DBpedia. In this case, we assume the URI to be the description of the predicate, i.e., $d_p = p$. We issue two different SPARQL queries to get outgoing

and incoming predicates, with respect to vertex v . The SPARQL query to retrieve d_p for an outgoing predicate p is:

```
QUERY: outgoingPredicate(v)
1 SELECT DISTINCT ?p
2 WHERE {
3   v ?p ?obj .}
```

Query incomingPredicate(v) is exactly the same, except from line 3, which becomes $?sub \ ?p \ v$. Both outgoing and incoming queries can be executed efficiently by most RDF engines that index all triples in six ways [59, 63] for traditional lookup. In some KGs, however, the predicate URIs might be arbitrary internal identifier¹². For example, a predicate with URI $p = \text{wdg:P227}$, where the description of wdg:P227 could also be retrieved by a SPARQL query from the KG wdg. After receiving the results of our queries, we check: If $?p$ is an arbitrary internal identifier, then we issue another query to get the predicate description.

Let r be a relation in $PGP(q)$, connecting two nodes $n^a, n^b \in PGP(q)$. The nodes are already linked to relevant vertices $\mathcal{R}_V(n^a, KG)$ and $\mathcal{R}_V(n^b, KG)$, respectively. Our goal is to link r . The intuition of our approach is that any successfully linked pair $\langle node, relation \rangle$ in the PGP must occur at least once in the target KG. Therefore, instead of blindly searching for predicates in KG , we limit our scope to those connected to vertices in $\mathcal{R}_V(n^a, KG) \cup \mathcal{R}_V(n^b, KG)$. For each such vertex v , we retrieve all its connected predicates by using SPARQL queries outgoingPredicate(v) and incomingPredicate(v). Note that we must check both directions because $PGP(q)$ is undirected.

For each retrieved predicate p and its corresponding description d_p , we use Equation 1 to compute an affinity score $S(l_r, d_p)$, where l_r is the label (see Definition 5.2) of r . Semantic match allows us to map, for instance, label “wife” to DBpedia predicate dbp:spouse. We select the top- k predicates, according to affinity score, as the set of relevant predicates for r . Formally:

Definition 5.2 (Relevant Predicates $\mathcal{R}_P$). Given a question q and a target knowledge graph $KG = (V, P)$ considered as a set of triples. Let r be a relation in $PGP(q)$ with label l_r and let $n^a, n^b \in PGP(q)$ be the nodes connected to r . Let $T_{rv} = \{v : \langle v, s_v \rangle \in \mathcal{R}_V(n^a, KG) \cup \mathcal{R}_V(n^b, KG)\}$ be the union of relevant vertices of n^a and n^b . The set of all outgoing and incoming predicates connected to relevant vertices is $T_{pd} = \{ \langle p, d_p, v, o \rangle : v \in T_{rv}, \langle v, p, ?obj \rangle \in KG \vee \langle ?sub, p, v \rangle \in KG \}$; flag o is *True* if v was an object in the triple, and labels d_p are retrieved as explained above. Let $T_p = \{ \langle p, S(l_r, d_p), v, o \rangle : \langle p, d_p, v, o \rangle \in T_{pd} \}$, where $S(o, o)$ is the semantic affinity score. The set of relevant predicates $\mathcal{R}_P(r, KG)$ for relation r in target KG , is the subset of T_p that contains all pairs with the top- k affinity scores.

For our running example q^E , the top-5 predicates for relation “city on shore”, assuming the target KG is DBpedia, are: dbp:city, dbp:locationCity, dbo:nearestCity, dbp:cities and dbp:country.

¹²Wikidata uses this naming system and provides a method to query the description, e.g., https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries#Adding_labels_for_properties

Algorithm 2 KGQANRelationLink

Input: r : a relation in $PGP(q)$, KG : target knowledge graph, k : number of predicates

Output: r annotated with relevant predicates $\mathcal{R}_P(r, KG)$

```

1:  $n^a, n^b \leftarrow PGP(q)$  nodes connected to  $r$ 
2:  $T_{rv} \leftarrow$  set of all  $v \in \mathcal{R}_V(n^a, KG) \cup \mathcal{R}_V(n^b, KG)$ 
3:  $T_{pd} \leftarrow \emptyset$  ▷ predicates of relevant vertices
4: for every  $v \in T_{rv}$  do
5:    $T_{pd} \leftarrow T_{pd} \cup \text{outgoingPredicate}(v)$  ▷ SPARQL to KG
6:    $T_{pd} \leftarrow T_{pd} \cup \text{incomingPredicate}(v)$ 
7: end for
8:  $T_p \leftarrow \emptyset$ 
9: for every  $\langle p, d_p, v, o \rangle \in T_{pd}$  do
10:  if not isHumanReadable( $p$ ) then
11:     $d_p \leftarrow \text{getPredicateDescription}(p)$ 
12:  end if
13:   $T_p \leftarrow T_p \cup \langle p, S(l_r, d_p), v, o \rangle$  ▷ compute semantic affinity
14: end for
15: return  $r.\mathcal{R}_P \leftarrow$  all pairs from  $T_p$  with top- $k$  affinity score

```

Algorithm 2 summarizes the relation linking process for a single relation r ; the algorithm must be called for every relation $r \in PGP(q)$. Line 2 constructs set T_{rv} that contains the union of relevant vertices of n^a and n^b , where n^a and n^b are the PGP nodes connected to r . In lines 3-7, for every relevant vertex v in T_{rv} , we execute at the target KG SPARQL queries `outgoingPredicate(v)` and `incomingPredicate(v)` to retrieve all predicates connected to v , together with their descriptions; we store the results into set T_{pd} . Then, lines 8-14 compute, for each predicate in T_{pd} , its semantic affinity to r , and store the resulting $\langle \text{predicate}, \text{score} \rangle$ pairs in T_p , refer to Definition 5.2. In lines 10-12, we check if p is not human-readable, e.g., a sequence of random characters and digits, then we issue a query to fetch the description of p . Finally, line 15 selects from T_p the pairs with the top- k score to construct the set of relevant predicates $\mathcal{R}_P(r, KG)$.

5.2.1 Complexity. Algorithm 2 is executed $|\mathcal{R}|$ times, where $\mathcal{R}$ is the set of relations in $PGP(q)$; see Definition 5.3. The most expensive part is the execution of the SPARQL queries. Each can be executed in constant time c_{lk} as a look-up query in the indices [59, 63] of the RDF engine. However, there are $2 \cdot |T_{rv}|$ such queries, where $|T_{rv}| \leq 2 \cdot \text{maxVR}$, since there are two nodes connected to r ; refer to the definition of `potentialRelevantVertices` query. In practice, we use only $k < \text{maxVR}$ vertices, therefore the cost of lines 4-7 becomes $O(k \cdot c_{lk})$. The cost of the for-loop in line 9 depends on T_{pd} , which can potentially contain all predicates from the target KG. However, in practice we limit the number of retrieved predicates per vertex, therefore the cost is $O(k)$; the same applies to line 15. The resulting complexity of Algorithm 2 is $O(k \cdot c_{lk} \cdot |\mathcal{R}|)$.

5.3 Annotated Graph Pattern (AGP)

Let q be a question with a corresponding graph $PGP(q)$, constructed during question understanding. Given a knowledge graph KG , Algorithms 1 and 2 generate, in a just-in-time manner, annotated graph

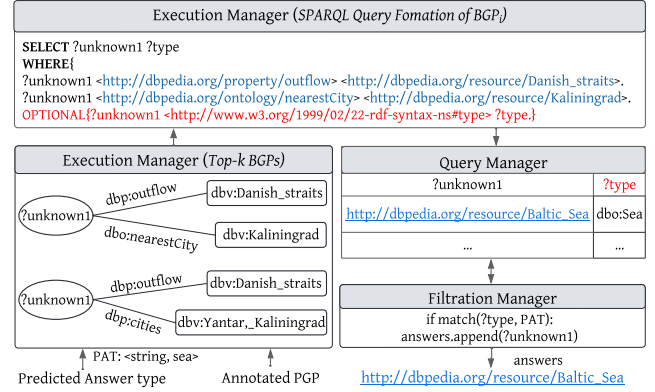


Figure 6: The KGQAN execution and Filtration phase uses the predicted answer type and the annotated PGP. KGQAN generates a set of BGPs and ranks them to create up to k queries. KGQAN adjusts these queries with an optional triple pattern to return the type of the main unknown for filtering.

$AGP(q, KG)$ that links the abstract components of $PGP(q)$ to actual vertices and predicates in KG , as shown in Figure 4. Formally:

Definition 5.3 (Annotated Graph Pattern (AGP)). Let $PGP(q) = (\mathcal{E}, \mathcal{R})$ be the phrase graph pattern for question q , and let KG be the target knowledge graph. The corresponding annotated graph pattern $AGP(q, KG)$ is an undirected graph consisting of nodes $\mathcal{E}$ and relations $\mathcal{R}$, where every $n \in \mathcal{E}$ is annotated with relevant vertices $\mathcal{R}_V(n, KG)$, and every $r \in \mathcal{R}$ is annotated with relevant predicates $\mathcal{R}_P(r, KG)$.

5.4 Semantic Affinity Calculation

Here we discuss how to compute the semantic affinity score $S(o, o)$ that we use during entity and relation linking. Let $l = (l_1, l_2, \dots)$ be a string consisting of a sequence of words. For each word $l_i \in l$, we generate a word embedding $E_w(l_i)$ using the FastText model [3], which is pre-trained on a large vocabulary of English words. The intuition is that, if two words are semantically similar, they will be close in the vector space of the embedding. If FastText cannot recognize l_i , then we generate a character embedding $E_c(l_i)$ instead, using the `chars2vec` [7, 8] pre-trained model, which captures the similarity of word spellings. Let l^X be a string and let $X = (x_1, x_2, \dots, x_{|l^X|})$ be an array of embeddings, where $x_i = E_w(l_i^X)$, if l_i^X appears in FastText, or $x_i = E_c(l_i^X)$, otherwise. Let l^Y also be a string, with its corresponding array of embedding Y . The semantic affinity [22] between l^X and l^Y , is defined as:

$$S(l^X, l^Y) = \frac{\sum_{x_i \in X, y_j \in Y} \text{sim}(x_i, y_j)}{|X| \cdot |Y|} \quad (1)$$

where $\text{sim}(o, o)$ is the cosine similarity. Equation 1 considers all pairs of (x_i, y_j) . Some of these pairs may contain embeddings from different models, for instance $(E_w(l_i^X), E_c(l_j^Y))$; in such cases, we define $\text{sim}(x_i, y_j) = 0$.

Algorithm 3 Generate the Top- k SPARQL queries from AGP

Input: $AGP(q, KG)$: an annotated graph pattern

Output: BGP_{SQ} : Top- k SPARQL queries

- 1: $BGP_{all} \leftarrow getBGPs(AGP(q, KG))$ $\triangleright$ all possible combinations
 - 2: $BGP_S \leftarrow calculateScores(BGP_{all})$ $\triangleright$ using equation 2
 - 3: $BGP_R \leftarrow rank(BGP_S)$
 - 4: $BGP_{SQ} \leftarrow getSPARQL(BGP_R, k)$ $\triangleright$ with optional type clause
-

While the above is our default semantic affinity calculation method, we also experiment with sentence-based embedding models. We use the GPT-3 [4] pre-trained transformer to generate a single embedding for the entire string. In this case, Equation 1 is simplified as: $S(l^X, l^Y) = sim(E_{GPT}(l^X), E_{GPT}(l^Y))$.

6 KGQAN EXECUTION AND FILTRATION

Our JIT approach shifts the filtration process from RDF engines to KGQAN and plans for it at runtime for a given question and an arbitrary KG. This shift helps KGQAN to work without pre-processing. Figure 6 illustrates the execution and filtration in KGQAN. The KGQAN post-filtering method percolates answers collected by executing the semantically equivalent queries (BGPs) against KG , which is a set of RDF triples $\langle sub, p, obj \rangle$. In KG , obj may be a certain class type for the vertex sub and p will be an $rdf:type^{13}$ predicate. KGQAN retrieves the class type of the main unknown, if available in KG , by extending the top selected BGPs by an optional triple pattern: $\langle unknown1, rdf:type, ?c \rangle$. Then, KGQAN generates a SPARQL query from this set of adjusted BGPs, sends them for execution to the RDF engine, and maintains the set $\{\langle a, ?c \rangle\}$, where a is a received answer and $?c$ is its associated class type. Finally, KGQAN iterates over this set to filter out each $\langle a, ?c \rangle$, where $?c$ does not match (i) the predicted data type, in case of date, or numerical; or (ii) the predicted semantic type, if the predicted data type is string.

Definition 6.1 (Basic Graph Pattern (BGP)). Let $AGP(q, KG) = (\mathcal{E}, \mathcal{R})$ be the annotated graph pattern for question q and target knowledge graph KG . The set of triple patterns in $AGP(q, KG)$ is $TP = \{\langle n^a, r, n^b \rangle : \langle n^a, r, n^b \rangle \in AGP(q, KG)\}$. For every triple in TP , assign a value $v^a \rightarrow n^a, v^b \rightarrow n^b$ and $p \rightarrow r$, such that $v^a \in n^a \cdot \mathcal{R}_V, v^b \in n^b \cdot \mathcal{R}_V$ and $p \in r \cdot \mathcal{R}_P$. The orientation $\langle v^a, p, v^b \rangle$, or $\langle v^b, p, v^a \rangle$ of the resulting triple, depends on flag o of p (see Definition 5.2).

Algorithm 3 explains KGQAN’s procedure for creating a list of ranked SPARQL queries from annotated graph pattern AGP . Line 1 creates BGP_{all} , a set of BGPs generated using all possible valid combinations of relevant vertices and predicates in AGP ; note that Definition 6.1 describes only *one* such BGP. In line 2, we calculate the score of each $BGP \in BGP_{all}$. The score of a BGP is defined as:

$$score(BGP) = \frac{1}{|TP|} \sum_{\langle v^a, p, v^b \rangle \in TP} (s_{v^a} + s_p + s_{v^b}) \quad (2)$$

where s_{v^a} , s_{v^b} , and s_p are the scores of the relevant vertices and predicates; refer to Definitions 5.1 and 5.2. Line 3 sorts BGP_S based on the calculated score. In line 4, we finally convert the top- k BGPs in

BGP_R to their equivalent SPARQL queries. Each query is appended by an optional clause with triple pattern $\langle unknown1, rdf:type, ?c \rangle$ to fetch the type or class linked to the main unknown, if any.

7 EXPERIMENTAL EVALUATION

7.1 Evaluation Setup

7.1.1 Compared Systems. We evaluate KGQAN against gAnswer [27, 64], EDGQA [28], and NSQA [31]. EDGQA is state-of-the-art in both QALD-9 and LC-QuAD 1.0. gAnswer was ranked first in the QALD-9 challenge [54]. NSQA utilized deep learning models to support question understanding and linking and outperformed gAnswer in QALD-9. The code of both gAnswer [25] and EDGQA [19] is available. We reproduce the results of gAnswer and EDGQA. NSQA uses a logical neural network [45], and special datasets for training the AMR model; both are not available. Thus, we only report the results provided at [31].

7.1.2 Four Different Real KGs. To evaluate these systems with diverse application domains, we use four real KGs, namely DBpedia, YAGO [16, 51], DBLP [11] and the Microsoft Academic Graph (MAG) [34]. Both DBpedia and YAGO are general-fact KGs, where most entities are about places and persons. DBLP and MAG are oriented to scientific publications, citations, authors and institutions, where some entities are identified by long phrases, such as a paper’s title and a conference name. Moreover, these four KGs are of different sizes to study the effect of the graph size on pre-processing required by these systems, as illustrated in Table 2.

7.1.3 Benchmarks and Question Sets. QA benchmarks include English questions annotated with the corresponding SPARQL queries and the set of correct answers from a specific version of a KG. QALD-9 [54] and LC-QuAD 1.0 [52] are widely used to evaluate QA systems on different versions of DBpedia. QALD-9 has 408 questions as a training set, and 150 as a testing set. The average length of a question in QALD-9 is 7.5 words. LC-QuAD 1.0 has a training set of 4000 questions and a testing set of 1000 questions. These questions are created based on different templates. There was no golden standard for YAGO, DBLP, or MAG. Thus, we asked a group of students with a background in computer science to express questions similar to QALD-9’s questions solved by most QA systems to find facts in YAGO, DBLP and MAG. We collected 100 questions per KG and annotated them with the correct SPARQL query and answers. These new benchmarks aim to evaluate the QA systems in solving similar questions to QALD-9 in unseen KGs and domains. Table 2 summarizes the five benchmarks we use. We used an evaluation metrics of: Precision (P), Recall (R), and Macro F1 (F1) and calculate them using the QALD-9 automatic evaluation tool [42].

7.1.4 RDF Engines for SPARQL Endpoints. We use Virtuoso 7.2.5.2 as SPARQL endpoints, as it is widely adopted as an endpoint for big KGs, such as DBpedia and MAG. We prepare five different Virtuoso endpoints for each of the benchmarks. The standard, unmodified installation of the Virtuoso engine was run at the endpoints and used by all QA systems in our experiments. The EDGQA linking method (Falcon) uses Elasticsearch. We use Elasticsearch 7.10.2 to index the KGs and enable the Falcon linking method to EDGQA.

¹³<https://www.w3.org/1999/02/22-rdf-syntax-ns#type>

Table 2: The used benchmarks, size of each KG, and time taken by the QA systems for pre-processing, i.e., indexing the KG.

Benchmarks	Benchmark Statistics			EDGQA - Indexing by Falcon		gAnswer	
	#Questions	KG Name	#Triples (M)	Index Time (hrs)	Index Size (G)	Index Time (hrs)	Index Size (G)
QALD-9	150	DBpedia-10	194	6.51	1.80	2.86	8.60
LC-QuAD 1.0	1000	DBpedia-04	140	6.23	1.70	2.28	6.60
YAGO-Bench	100	YAGO-4	145	6.88	2.00	1.81	4.10
DBLP-Bench	100	DBLP	136	4.83	1.60	1.91	5.20
MAG-Bench	100	MAG	13000	103.22	92.00	37.40	319.00

7.1.5 Computing Infrastructure. We deploy each of the evaluated QA systems plus KGs on virtuoso SPARQL endpoints running on the same local machine. We use two different settings for our experiments. In the first setting, we use two Linux machines, each with 16 cores and 500GB RAM. These machines are used with all experiments except those related to MAG. In the second setting, we use five Linux machines with 32 cores and 3TB RAM. EDGQA and gAnswer use these five machines to pre-process MAG and generate the necessary indices for their approach to work.

7.1.6 KGQAN implementation and settings. KGQAN is implemented¹⁴ using Python 3.7 and all models are trained using Pytorch 1.11.0. Our implementation supports parallel execution and filtration via multi-threading. We did not enable multi-threading for the execution and filtration, as gAnswer and EDGQA do not have parallel support. We utilize BART [32] for modelling our QU Seq2Seq model. Our semantic affinity model uses FastText’s wiki-news-300d-1M model. We also show experiments with GPT-3 [4] for our QU and semantic affinity models in Sub-section 7.3. KGQAN needs four parameters: (i) *Max Fetched Vertices* to decide how many vertices for a certain entity could be investigated, (ii) *Number of Vertices* to decide how many vertices could be used to annotate each node in the PGP, (iii) *Number of Predicates* to decide how many predicates could be used to annotate each edge in the PGP, and (iv) *Max number of Queries* to decide the number of equivalent SPARQL queries to be generated for the question. Our experiments use the following values 400, 1, 20, and 40, respectively. We decide the number of predicates based on the general average number of predicates per vertex. We tuned these parameters to work across different KGs, *not to outperform for a specific one*.

7.2 Experiments with Real KGs

gAnswer, EDGQA, and NSQA are trained and evaluated using QALD-9 and LC-QuAD 1.0. We compare KGQAN to these systems using five benchmarks on diverse real KGs. We performed the pre-processing required by gAnswer and EDGQA. Table 2 summarizes the time and storage consumed by each system for pre-processing. We assessed these systems with seen, i.e., QALD-9 and LC-QuAD 1.0, and unseen benchmarks on YAGO, DBLP and MAG. Table 3 shows the results of all systems on the five benchmarks.

7.2.1 Pre-processing Cost. EDGQA uses an ensemble of three different linking methods, namely Falcon [47], Dexter [6], and EARL[18] for entity and relation linking. We use Falcon to index all KGs and enable EDGQA to work. To reproduce EDGQA’s results in QALD-9

and LC-QuAD 1.0 we use the indices for Dexter and EARL, which are provided at the EDGQA repository. For MAG, we customize Falcon’s code with the right predicate describing each entity type, such as the paper’s title for papers and author’s name for authors. gAnswer provided their indexing mechanism and relation mentions file that are needed in the pre-processing. We use the machines with 3TB of RAM to run the gAnswer indexing script. For EDGQA, Falcon consumes more time in indexing a KG than gAnswer’s indexing method, as shown in Table 2.

7.2.2 Seen Benchmarks. For the LC-QuAD 1.0 benchmark, KGQAN performs almost identically to EDGQA in F1 score and an improvement of 15% in precision, as shown Table 3. EDGQA outperformed KGQAN in the recall by almost 19% due to the aggressive indexing of three different linking systems and the human-curated rules to understand questions of LC-QuAD 1.0. These questions are automatically created using templates, i.e., curating rules per template helps perform well in training and testing questions. Unlike LC-QuAD 1.0, QALD-9’s questions are created manually with different complexity. Thus, QALD-9 is more challenging, e.g., curated rules may work with training questions and fail with testing questions. This explains the huge difference in EDGQA’s F1 scores in LC-QuAD 1.0 and QALD-9. KGQAN outperforms the existing QA systems in QALD-9, as shown Table 3, by achieving an impressive precision and F1 score of 49.81 and 43.99, respectively.

7.2.3 Unseen Benchmarks on YAGO, DBLP and MAG. To evaluate the universality of the systems in processing a question against an arbitrary KG, we use three benchmarks unseen by all systems, including KGQAN. One benchmark has questions similar to QALD-9 and targets YAGO, a KG similar to DBpedia. Another two benchmarks target a domain different from DBpedia, where the questions find facts related to papers and authors in DBLP and MAG. Interestingly, gAnswer achieved better performance in the YAGO benchmark than the LC-QuAD 1.0 benchmark, as the questions in YAGO are similar to the ones in QALD-9. In MAG and most cases in DBLP, vertices’ URIs end with code, e.g., “Jim Gray”’s URI is `mag_e:2279569217`, see Subsection 5.1. The gAnswer inverted index is based on the URIs. So, gAnswer linking cannot find “Jim Gray”. Thus, gAnswer answers two questions in DBLP and Zero in MAG due to its QU and linking. On DBLP and MAG, EDGQA fails to answer most questions, as it failed to understand questions about entities with longer phrases. KGQAN’s question understanding model generalizes better in extracting entities with longer phrases. Moreover, KGQAN’s semantic affinity model is trained based on general English text. This helps our JIT linking work well across KGs of different domains and utilize the RDF engines’ built-in indices.

¹⁴<https://github.com/CoDS-GCS/KGQAN>

Table 3: Results of five benchmarks. KGQAN’s recall and precision across KGs of different domains proves that KGQAN performs better than other systems due to our Seq2Seq model and JIT linking that work without a pre-processing phase for a specific KG.

System	QALD-9-DBpedia			LC-QuAD 1.0-DBpedia			YAGO			DBLP			MAG		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
NSQA	31.89	32.05	31.26	44.76	45.82	44.45	-	-	-	-	-	-	-	-	-
gAnswer	29.34	32.68	29.81	82.21	4.31	8.18	58.49	34.05	43.04	78.00	2.00	3.90	0.0	0.0	0.0
EDGQA	31.30	40.30	32.00	50.50	56.00	53.10	41.90	40.80	41.40	8.00	8.00	8.00	4.00	4.00	4.00
KGQAN	51.13	38.72	44.07	58.71	46.11	51.65	48.48	65.22	55.62	57.87	52.02	54.79	55.43	45.61	50.05

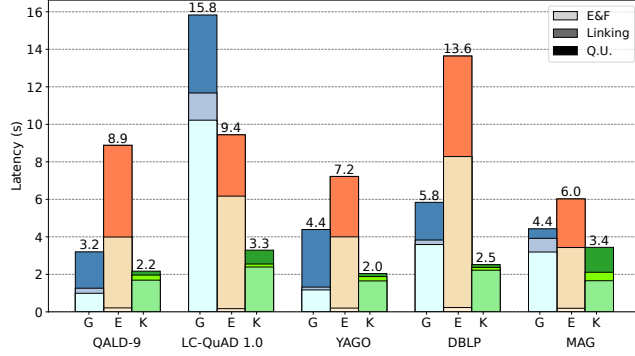


Figure 7: Response time of gAnswer (G), EDGQA (E) and KGQAN (K). Each bar shows average response time classified bottom-up into QU, Linking, and Execution/Filteration (E&F).

7.2.4 Response time. This experiment analyzes the QA systems in terms of response time to a question. Per the system, we calculate the average response time for every set of questions. We use a local setting where the QA system and the benchmark SPARQL endpoint are deployed in the same machine to eliminate variability due to network latency or workload on the actual endpoint of KGs, such as MAG or DBpedia. We report the average time of each step, question understanding (QU), linking, and execution/filteration, as shown in Figure 7. The KGQAN QU model consumes the majority of response time. The KGQAN linking usually consumes the lowest time. Finally, the execution and filtration in KGQAN consume more time than linking, depending on how complex the generated SPARQL queries are and the size of the query results.

In gAnswer and EDGQA, the total response time is dominated by the accuracy of question understanding in terms of: (i) the extracted entities and relations, which affect the linking time, and (ii) the number of extracted triples, which affects the number of generated SPARQL queries, i.e., execution time. The total response time is dominated by the complexity of the overall question-answering pipeline more than the graph size. For example, KGQAN consumes similar time in LC-QuAD 1.0 and MAG. In general, EDGQA consumes more time in linking as the utilized linking methods use disk-based indices. In contrast, gAnswer loads its indices in memory before processing any question.

7.3 Analyzing KGQAN

7.3.1 Question Understanding. This experiment analyzes the QA systems in terms of total failure in answering a question, i.e., recall and F1 are zeros. This failure could be due to QU or other reasons,

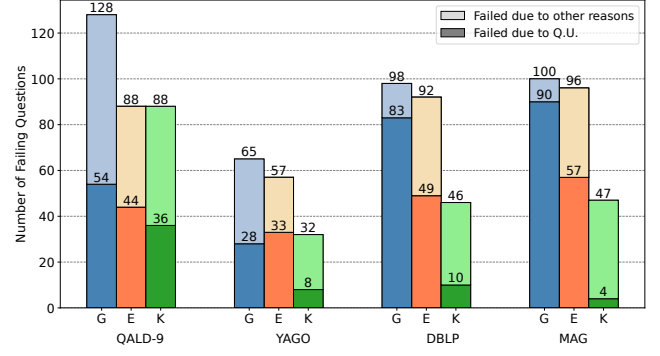


Figure 8: Number of failing questions with R=0 and F1=0 in each benchmark. Each bar shows the total number classified bottom-up into failing due to QU or others. KGQAN fails in the least number of questions across diverse benchmarks.

such as linking or filtering. Failing to understand a question means no entities or relation phrases are correctly detected. We manually count the failing cases due to QU in each benchmark, as summarized in Figure 8. KGQAN outperforms gAnswer and EDGQA in understanding most of the questions across benchmarks of different domains, as it fails in the least number of questions in total and due to QU per benchmark. KGQAN was able to understand questions in an unseen domain, i.e., DBLP, better than gAnswer and EDGQA.

7.3.2 JIT Linking and Post-Filtering. A labeled dataset for the entity and relation linking task in LC-QuAD 1.0 is provided by [18]. In this experiment, we use this dataset to analyze the performance of the linking methods adopted by gAnswer and EDGQA, as shown in Figure 9. Our JIT linking approach eliminates the need for the pre-processing phase by offloading the linking task partially as queries executed by the RDF engine and followed by semantic ranking at KGQAN. EDGQA utilizes three different linking systems. Thus, it achieves an outstanding performance in this task. gAnswer’s QU is trained only on QALD-9. Therefore, it does not extract the correct entities and relations in most cases, and consequently, it does not perform well in the linking task in LC-QuAD 1.0. Our approach aims at maximizing the recall to find the right vertex. KGQAN performs a post-filtration of irrelevant answers to improve the precision. Hence, the final F1 score of KGQAN is almost identical to the highest F1 score achieved in the entity linking. In contrast, the overall processing in EDGQA negatively affects the F1 score achieved by the three linking systems.

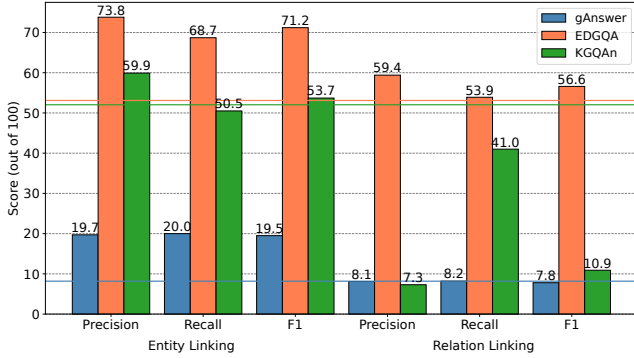


Figure 9: The entity and relation linking using LC-QuAD 1.0 show that EDGQA benefits from combining three linking methods. The horizontal lines show the final F1 score per system. Unlike EDGQA, KGQAN achieves a final F1 score almost identical to its highest F1 in the entity linking.

Table 4: KGQAN’s performance using different pre-trained models for training our QU and semantic affinity (SA) models. Our default settings using BART and fine-grained (FG) affinity achieve better F1 scores in most cases than the variations with GPT-3 in both QU and the coarse-grained (CG) affinity.

Benchmarks	QU: BART	QU: GPT-3	QU: BART
	SA: FG	SA: FG	SA: GPT-3
QALD-9	44.07	42.12	42.60
LC-QuAD 1.0	51.65	52.87	50.86
YAGO	55.62	54.94	55.02
DBLP	54.79	54.42	41.72
MAG	50.05	49.26	37.64

7.3.3 Filtration effect. This experiment analyzes KGQAN performance with and without filtration. For the lack of space, we show only our results on QALD-9 and LC-QuAD 1.0, as illustrated in Figure 10. KGQAN predicts answer data type. Our model performs very well in filtering answers based on the data types date, numerical or boolean. Due to the high ambiguity, filtering answers using *semantic* types is not as accurate as the data types date, numerical or boolean. Overall, our filtration method is designed to avoid hurting the recall much. QALD-9 has a higher ratio of questions whose answers are of type date, numerical or boolean than LC-QuAD 1.0. Thus, KGQAN with filtration achieved better in QALD-9 than LC-QuAD 1.0.

7.3.4 KGQAN and different pre-trained language models. This experiment analyzes the effectiveness of different pre-trained language models (PLMs) on the KGQAN performance in both question understanding and semantic affinity, which we use in both linking and filtration. For question understanding, we utilize BART [32] and GPT-3 [4] to model our triple patterns extraction task. For BART, we use the Huggingface API [29] to perform the training, as it gives us the freedom to fine-tune the training parameters. Fine-tuning the GPT-3 model is only available through an OpenAI API [37]. Thus, we had less control in optimizing our model using GPT-3.

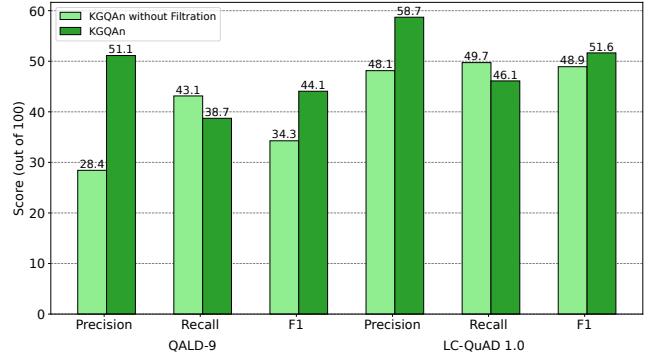


Figure 10: Analyzing KGQAN’s performance with and without filtering, using QALD-9 and LC-QuAD 1.0. Filtering improves precision but slightly reduces recall. The resulting F1 score is improved for both benchmarks.

Our semantic affinity model depends on the representation (embeddings) of two sets of words. We develop a fine-grained approach to estimate the affinity at the granularity of a pair of words. For this approach, we use FastText and chars2vec models. As a coarse-grained approach, we use GPT-3 to get one vector embedding for each set of words.

We compare the performance of KGQAN using these different combinations of models. KGQAN by default utilizes our BART-based Seq2Seq and the fine-grained semantic affinity (FG) models. We create two different variations of KGQAN where we replace our BART-based Seq2Seq model in one variation with our model trained using GPT-3 for question understanding. In the second variation, we use our BART-based Seq2Seq model and the coarse-grained semantic affinity (CG) based on GPT-3. Table 4 illustrates the final F1 score for each benchmark using KGQAN based on the default setting, and the two variations, from left to right. Our default setting outperforms the two variations in most cases.

7.4 Taxonomy of benchmark questions

We develop a taxonomy to study the complexity of the benchmark questions based on different characteristics, and evaluate the performance of KGQAN, gAnswer and EDGQA based on this taxonomy. As illustrated in Table 5, our taxonomy categorizes the complexity of the questions based on (i) the shape of the candidate SPARQL query and (ii) the linguistic complexity. A SPARQL query could be classified into a star or path query [39]. In our taxonomy, a star query is a query consisting of one or more triple patterns sharing the same subject. A path query contains at least two triple patterns where an object in one triple pattern is a subject in another pattern, i.e., a linear shape. Based on the linguistic classification provided by LC-QuAD 2.0 [17], the questions in QALD-9 and the unseen benchmarks are classified into single fact, single fact with type, multi-facts, and booleans. KGQAN outperforms gAnswer and EDGQA in most cases across the different benchmarks.

Table 5: Number of questions solved by the three systems compared using the taxonomy of the benchmark questions which depends on the SPARQL query shapes and the linguistic complexity of the questions according to LC-QuAD 2.0 [17]

Query type	SPARQL shape								LC-QuAD 2.0 taxonomy															
	Star				Path				Single fact				Fact with type				Multi fact				Boolean			
Benchmark	# queries	KGQAn	EDGQA	gAnswer	# queries	KGQAn	EDGQA	gAnswer	# queries	KGQAn	EDGQA	gAnswer	# queries	KGQAn	EDGQA	gAnswer	# queries	KGQAn	EDGQA	gAnswer	# queries	KGQAn	EDGQA	gAnswer
QALD-9	131	60	56	21	19	2	5	0	81	46	41	16	28	7	8	3	37	9	9	2	4	0	3	0
YAGO-B	92	63	39	32	8	5	4	3	87	61	40	33	6	5	3	2	6	2	0	0	1	0	0	0
DBLP-B	92	46	8	2	8	8	0	0	85	49	8	1	11	4	0	1	4	1	0	0	0	-	-	-
MAG-B	77	44	4	0	23	9	0	0	75	40	4	0	7	2	0	0	16	9	0	0	2	2	0	0

8 RELATED WORK

There is a growing effort to develop QA systems based on different techniques for question understanding and linking [38, 43]. QA systems are classified into SPARQL- and non-SPARQL-based. Examples of non-SPARQL-based systems are QAMP [55], Treo [23], and others [60, 62]. These systems do not translate questions into SPARQL queries. Instead, they implement proprietary query engines; therefore, they cannot be used with arbitrary SPARQL endpoints.

SPARQL-based systems translate natural language questions into SPARQL queries using various techniques for: (i) question understanding, to generate an intermediate abstract representation; and (ii) linking the abstract representation to vertices and predicates of a KG. Examples include gAnswer [27, 64], EDGQA [28], NSQA [31], WDAqua-core1 [15], Bio-SODA [48], and QAnswerKG [14].

The question understanding techniques are classified into: (i) rule-based, such as gAnswer [27, 64], EDGQA [28], and WDAqua-core1 [15]. They curate rules based on Part-of-Speech (POS) tagging to extract entities and relation phrases. Unlike KGQAN, systems based on human-curated rules are hard to generalize to a broad set of applications and groups of diverse users. (ii) deep learning-based, such as NSQA [31], which curates tree representations of a large set of questions and uses deep learning to train a semantic parser called AMR. Note that, AMR has different granularity than SPARQL; thus, there is a need to adjust NSQA to work with new domains [31]. KGQAN does not suffer this issue, as SPARQL queries are generated by traversing the PGP, whose nodes and edges are annotated by relevant vertices and predicates in a KG. (iii) token-based techniques [14, 48] that extract the longest sequence of keywords in a question that matches an entry in an inverted index built by these systems. Token-based systems are KG specific.

KGQAN annotates the generated PGP using our just-in-time approach that does not demand any pre-processing of the target KG. Unlike KGQAN, existing systems index the target KG in advance, such as EDGQA [28] and gAnswer [27, 64], or build a vector space of all KG’s vertices and predicates to predict the linking based on deep learning models, such as NSQA [31]. In these systems, the pre-processing cost is proportional to the KG’s size. For example, in gAnswer, the pre-processing complexity is polynomial to the number of KG’s vertices [27, 64]. EDGQA utilizes three different systems to index the target KG, where each system indexes the KG differently.

KGQAN extracts a sequence of triple patterns from a question; entities can be unknowns (i.e., variables). Unlike our problem, existing relation triple extraction techniques [9, 58] do not deal with unknowns; furthermore, relations are chosen from a pre-defined list, that is, they are not extracted from the given text. Thus, relation triple extraction models and their training datasets cannot be used for our task.

We train our task as a Seq2Seq model using BART [32]. Existing systems also use Seq2Seq models. For example, Ref. [46] is trained with extracted KG-specific information to generate a sequence of predicates for relations in a question. Also, Ref. [57] gets an annotated question and generates an annotated SQL query. QAMP [55] also utilizes a Seq2Seq model, not as a generation task, but to label sequences in a question as an entity, predicate, or class. Unlike these models, KGQAN’s Seq2Seq model does not depend on a query language, a specific KG, or a particular domain. Thus, our model has more flexibility to understand questions in different domains without needing a domain expert to prepare a training set.

9 CONCLUSION

Existing QA systems are domain- and KG-specific and demand an expensive pre-processing phase. Thus, they cannot be used on-demand to process a question against an arbitrary KG. This paper presents KGQAN, an on-demand KG question-answering platform that overcomes these limitations. KGQAN proposes a novel formalization of question understanding as a triple pattern extraction modelled using a Seq2Seq neural network. Our model generalizes to understand questions across diverse domains. Moreover, KGQAN introduces a just-in-time linking and filtering approach, which performs entity and relation linking as semantic search queries partially offloaded to the RDF engines. We evaluate the state-of-the-art (SOTA) QA systems using broadly utilized benchmarks and four diverse real-life KGs. Based on the KG size, KGQAN saves a few hours to days of pre-processing. KGQAN achieves comparable F1 score to the SOTA for the common benchmarks, but outperforms the SOTA by a large margin for previously unseen KGs. In our future work, we plan to expand our question understanding training to include more complex questions; we also plan to support multi-intention questions.

Acknowledgement. We thank the authors of EDGQA, gAnswer, and Falcon for their assistance to reproduce their results.

REFERENCES

- [1] Ibrahim Abdelaziz, Essam Mansour, Mourad Ouzzani, Ashraf Aboulhag, and Panos Kalnis. 2017. Lusal: A System for Querying Linked Data at Scale. *Proceedings of the VLDB Endowment*, (PVLDB) 11, 4 (2017), 485–498. <http://www.vldb.org/pvldb/vol11/p485-abdelaziz.pdf>
- [2] Waqas Ali, Muhammad Saleem, Bin Yao, Aidan Hogan, and Axel-Cyrille Ngonga Ngomo. 2022. A survey of RDF stores & SPARQL engines for querying knowledge graphs. *VLDB Journal* 31, 3 (2022), 1–26. <https://doi.org/10.1007/s00778-021-00711-3>
- [3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguistics* (2017), 135–146. <https://transaclog.org/ojs/index.php/tacl/article/view/999>
- [4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc9467418bfb8ac142f64a-Abstract.html>
- [5] Damian Bursztyjn, François Goasdoué, Ioana Manolescu, and Alexandra Roatis. 2015. Reasoning on web data: Algorithms and performance. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, Vol. 1272. 1541–1544. <https://doi.org/10.1109/ICDE.2015.7113422>
- [6] Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2014. Dexter 2.0 - an Open Source Tool for Semantically Enriching Data. In *Proceedings of the International Semantic Web Conference, Posters & Demonstrations Track (ISWC)*, Vol. 1272. 417–420. http://ceur-ws.org/Vol-1272/paper_127.pdf
- [7] Chars2vec: Character-based word embeddings model based on rnn, Article. 2019. <https://hackernoon.com/chars2vec-character-based-language-model-for-handling-real-world-texts-with-spelling-errors-and-a3e4053a147d>
- [8] Chars2vec: Character-based word embeddings model based on rnn, Github. 2019. <https://github.com/IntuitionEngineeringTeam/chars2vec>
- [9] Yubo Chen, Yunqi Zhang, and Yongfeng Huang. 2022. Learning Reasoning Patterns for Relational Triple Extraction with Mutual Generation of Text and Graph. In *Findings of the Association for Computational Linguistics (ACL)*. 1638–1647. <https://aclanthology.org/2022.findings-acl.129>
- [10] CoreNLP. 2022. <https://stanfordnlp.github.io/CoreNLP/>
- [11] DBLP release. 2022. <https://dblp.org/rdf/release/dblp-2022-06-01.nt.gz>
- [12] Marie-Catherine De Marneffe and Christopher D Manning. 2016. *Stanford typed dependencies manual*. Technical Report. Stanford University. https://nlp.stanford.edu/software/dependencies_manual.pdf
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*. 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [14] Dennis Diefenbach, José Giménez-García, Andreas Both, Kamal Singh, and Pierre Maret. 2020. QAnswer KG: Designing a Portable Question Answering System over RDF Data. In *Proceedings of the International Conference of the Semantic Web (ESWC)*, Vol. 12123. 429–445. https://doi.org/10.1007/978-3-030-49461-2_25
- [15] Dennis Diefenbach, Kamal Deep Singh, and Pierre Maret. 2018. WDAqua-core1: A Question Answering service for RDF Knowledge Bases. In *Companion Proceedings of the The Web Conference, (ACM)*. 1087–1091. <https://doi.org/10.1145/3184558.3191541>
- [16] YAGO downloads. 2022. <https://yago-knowledge.org/downloads/yago-4>
- [17] Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In *Proceedings of The Semantic Web - (ISWC)*, Vol. 11779. 69–78. https://doi.org/10.1007/978-3-030-30796-7_5
- [18] Mohnish Dubey, Debayan Banerjee, Debanjan Chaudhuri, and Jens Lehmann. 2018. EARL: Joint Entity and Relation Linking for Question Answering over Knowledge Graphs. In *Proceedings of the International Semantic Web Conference (ISWC)*, Vol. 11136. 108–126. https://doi.org/10.1007/978-3-030-00671-6_7
- [19] EDGQA code. 2021. <https://github.com/HXX97/EDGQA>
- [20] EDGQA Log. 2021. https://raw.githubusercontent.com/HXX97/EDGQA/main/query_logs/query_log_test_lc-quad_2021_07_08%2023_48.txt
- [21] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1535–1545. <https://aclanthology.org/D11-1142/>
- [22] Raul Fernandez, Essam Mansour, Abdulhakim Qahtan, Ahmed Elmagarmid, and Ihab Ilyas et al. 2018. Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery. In *Proceedings of the International Conference on Data Engineering, (ICDE)*. 989–1000. <https://doi.org/10.1109/ICDE.2018.00093>
- [23] André Freitas, João Gabriel Oliveira, Seán O’Riain, João Carlos Pereira da Silva, and Edward Curry. 2013. Querying linked data graphs using semantic relatedness: A vocabulary independent approach. *Data and Knowledge Engineering (DKE)* 88 (2013), 126–141. <https://doi.org/10.1016/j.datak.2013.08.003>
- [24] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the International World Wide Web Conference (WWW)*. 413–422. <https://doi.org/10.1145/2488388.2488425>
- [25] GAnswer code. 2022. <https://github.com/pkumod/gAnswer>
- [26] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, and et al. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. 1–6. <https://aclanthology.org/W18-2501>
- [27] Sen Hu, Lei Zou, Jeffrey Yu, Haixun Wang, and Dongyan Zhao. 2018. Answering Natural Language Questions by Subgraph Matching over Knowledge Graphs. *IEEE Transactions on Knowledge and Data Engineering, (TKDE)* 30 (2018), 824–837. <https://doi.org/10.1109/TKDE.2017.2766634>
- [28] Xixin Hu, Yiheng Shu, Xiang Huang, and Yuzhong Qu. 2021. EDG-Based Question Decomposition for Complex Question Answering over Knowledge Bases. In *Proceedings of the International Semantic Web Conference, (ISWC)*. 128–145. https://doi.org/10.1007/978-3-030-88361-4_8
- [29] Huggingface. 2022. <https://huggingface.co/>
- [30] Jena text search. 2021. <https://jena.apache.org/documentation/query/text-query.html>
- [31] Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, and Alexander G. Gray et al. 2021. Leveraging Abstract Meaning Representation for Knowledge Base Question Answering. In *Findings of the Association for Computational Linguistics: (ACL/IJCNLP)*. 3884–3894. <https://doi.org/10.18653/v1/2021.findings-acl.339>
- [32] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, and Abdelrahman Mohamed et al. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, (ACL)*. 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
- [33] Xueling Lin, Haoyang Li, Hao Xin, Zijian Li, and Lei Chen. 2020. KBPearl: A Knowledge Base Population System Supported by Joint Entity and Relation Linking. *Proceedings VLDB Endowment (PVLDB)* 13, 7 (2020), 1035–1049. <https://doi.org/10.14778/3384345.3384352>
- [34] MAG records. 2022. <https://zenodo.org/record/4617285?YrNszNLMJhH>
- [35] Ndapandula Nakashole, Gerhard Weikum, and Fabian M. Suchanek. 2012. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. 1135–1145. <https://aclanthology.org/D12-1104/>
- [36] Tahirah Naseem, Srinivas Ravishankar, Nandana Mihindukulasooriya, Ibrahim Abdelaziz, and Young-Suk Lee. 2021. A Semantics-aware Transformer Model of Relation Linking for Knowledge Base Question Answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, (ACL/IJCNLP)*. 256–262. <https://doi.org/10.18653/v1/2021.acl-short.34>
- [37] OpenAI. 2022. <https://openai.com/>
- [38] Fatma Özcan, Abdul Quamar, Jaydeep Sen, Chuan Lei, and Vasilis Efthymiou. 2020. State of the Art and Open Challenges in Natural Language Interfaces to Data. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. ACM, 2629–2636. <https://doi.org/10.1145/3318464.3383128>
- [39] M. Tamer Özsu. 2016. A survey of RDF data management systems. In *proceedings of Frontiers Comput. Sci.* 10, 3 (2016), 418–432. <https://doi.org/10.1007/s11704-016-5554-y>
- [40] Jeff Pan, Mei Zhang, Kuldeep Singh, Frank Harmelen, and Jinguang Gu et al. 2019. Entity Enabled Relation Linking. In *Proceedings of International Semantic Web Conference, (ISWC)*, Vol. 11778. 523–538. https://doi.org/10.1007/978-3-030-30793-6_30
- [41] Predefined Synonyms. 2022. https://drive.google.com/file/d/1hmqafrTo0_qQNRAPcuxFXaBx7S0sNVy/view
- [42] QALD scripts. 2022. <https://github.com/ag-sc/QALD/blob/master/6/scripts/evaluation.rb>
- [43] Abdul Quamar, Vasilis Efthymiou, Chuan Lei, and Fatma Özcan. 2022. Natural Language Interfaces to Data. *Found. Trends Databases* 11, 4 (2022), 319–414. <https://doi.org/10.1561/19000000078>
- [44] Hoan Nguyen Mau Quoc, M Serrano, HN Mau, JG Breslin, and D Le-Phuoc. 2019. A performance study of RDF stores for linked sensor data. *Semantic Web Journal* 1 (2019), 72. <http://www.semantic-web-journal.net/system/files/swj2249.pdf>
- [45] Ryan Riegel, Alexander G. Gray, Francois P. S. Luus, Naweid Khan, and Ndivhuwo Makondo et al. 2020. Logical Neural Networks. *CoRR abs/2006.13155* (2020). [arXiv:2006.13155](https://arxiv.org/abs/2006.13155) <https://arxiv.org/abs/2006.13155>
- [46] Gaetano Rossiello, Nandana Mihindukulasooriya, Ibrahim Abdelaziz, Mihaela A. Bornea, and Alfio Gliozzo et al. 2021. Generative Relation Linking for Question Answering over Knowledge Bases. In *Proceedings of the Semantic Web Conference (ISWC) (Lecture Notes in Computer Science)*, Vol. 12922. Springer, 321–337. https://doi.org/10.1007/978-3-030-88361-4_19
- [47] Ahmad Sakor, Isaiah Mulang, Kuldeep Singh, Saeedeh Shekarpour, and Maria-Esther Vidal et al. 2019. Old is Gold: Linguistic Driven Approach for Entity and

- Relation Linking of Short Text. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 2336–2346. <https://doi.org/10.18653/v1/n19-1243>
- [48] Ana Sima, Tarcisio Farias, Maria Anisimova, Christophe Dessimoz, and Marc Rechavi et al. 2021. Bio-SODA: Enabling Natural Language Question Answering over Knowledge Graphs without Training Data. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*. 61–72. <https://doi.org/10.1145/3468791.3469119>
- [49] Valentin I. Spitskovsky and Angel X. Chang. 2012. A Cross-Lingual Dictionary for English Wikipedia Concepts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 3168–3175. <http://www.lrec-conf.org/proceedings/lrec2012/summaries/266.html>
- [50] Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. 3104–3112. <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>
- [51] Thomas Tanon, Gerhard Weikum, and Fabian Suchanek. 2020. YAGO 4: A Reasonable Knowledge Base. In *Proceedings of the European Semantic Web Conference (ESWC)*, Vol. 12123. 583–596. https://doi.org/10.1007/978-3-030-49461-2_34
- [52] Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs. In *Proceedings of the International Semantic Web Conference (ISWC)*, Vol. 10588. 210–218. https://doi.org/10.1007/978-3-319-68204-4_22
- [53] Immanuel Trummer. 2022. From BERT to GPT-3 Codex: Harnessing the Potential of Very Large Language Models for Data Management. *Proc. VLDB Endow.* 15, 12 (2022), 3770–3773. <https://www.vldb.org/pvldb/vol15/p3770-trummer.pdf>
- [54] Ricardo Usbeck, Ria Gusmita, Axel-Cyrille Ngomo, and Muhammad Saleem. 2018. 9th Challenge on Question Answering over Linked Data (QALD-9). In *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)*, Vol. 2241. 58–64. <http://ceur-ws.org/Vol-2241/paper-06.pdf>
- [55] Svitlana Vakulenko, Javier Garcia, Axel Polleres, Maarten Rijke, and Michael Cochez. 2019. Message Passing for Complex Question Answering over Knowledge Graphs. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 1431–1440. <https://doi.org/10.1145/3357384.3358026>
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, and Llion Jones et al. 2017. Attention is All you Need. In *Proceedings of the Advances in neural information processing systems (NeurIPS)*. 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [57] Wenlu Wang, Yingtao Tian, Haixun Wang, and Wei-Shinn Ku. 2020. A Natural Language Interface for Database: Achieving Transfer-learnability Using Adversarial Method for Question Understanding. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*. IEEE, 97–108. <https://doi.org/10.1109/ICDE48307.2020.00016>
- [58] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 1476–1488. <https://doi.org/10.18653/v1/2020.acl-main.136>
- [59] Cathrin Weiss, Panagiotis Karras, and Abraham Bernstein. 2008. Hexastore: sextuple indexing for semantic web data management. *Proceedings of VLDB Endowment (PVLDB)* 1, 1 (2008). <https://doi.org/10.14778/1453856.1453965>
- [60] Dekun Wu, Nana Nosirova, Hui Jiang, and Mingbin Xu. 2019. A General FOFE-net Framework for Simple and Effective Question Answering over Knowledge Bases. *Computing Research Repository, (CoRR)* abs/1903.12356 (2019). [arXiv:1903.12356](http://arxiv.org/abs/1903.12356)
- [61] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6397–6407. <https://doi.org/10.18653/v1/2020.emnlp-main.519>
- [62] Xuchen Yao and Benjamin Durme. 2014. Information Extraction over Structured Data: Question Answering with Freebase. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 956–966. <https://doi.org/10.3115/v1/p14-1090>
- [63] Pingpeng Yuan, Pu Liu, Buwen Wu, Hai Jin, Wenya Zhang, and Ling Liu. 2013. TripleBit: a Fast and Compact System for Large Scale RDF Data. *Proceedings of VLDB Endowment (PVLDB)* 6, 7 (2013), 517–528. <https://doi.org/10.14778/2536349.2536352>
- [64] Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Yu, and Wenqiang He et al. 2014. Natural language question answering over RDF: a graph data driven approach. In *Proceedings of the International Conference on Management of Data (SIGMOD)*. 313–324. <https://doi.org/10.1145/2588555.2610525>