Schema and content aware classification for predicting the sources containing an answer over corpus and knowledge graphs

Somayeh Asadifar [1] , **Mohsen Kahani** [Corresp., 1] , **Saeedeh Shekarpour** [2]

[1] Faculty of engineering, Ferdowsi University of Mashhad, Mashhad, Khorasan razavi, Iran

[2] College of Arts and Sciences: Computer Science, University of Dayton, Dyton, Ohio, United States

Corresponding Author: Mohsen Kahani
Email address: kahani@um.ac.ir

Today, several attempts to manage question answering (QA) have been made in three separate areas: 1) Knowledge-based (KB), 2) Text-based and 3) Hybrid, which takes advantage of both prior areas in extracting the response. On the other hand, in question answering on a large number of sources, source prediction to ensure scalability is very important. In this paper, a method for source prediction is presented in hybrid QA, involving several KB sources and a text source. In a few hybrid methods for source selection, including only one KB source in addition to the textual source, prioritization or heuristics have been used that have not been evaluated so far. Most methods available in source selection services are based on general metadata or triple instances. These methods are not suitable due to the unstructured source in hybrid QA. In this research, we need data details to predict the source. In addition, unlike KB federated methods that are based on triple instances, we use the behind idea of mediated schema to ensure data integration and scalability. Results from evaluations that consider word, triple, and question level information, show that the proposed approach performs well against a few benchmarks. In addition, the comparison of the proposed method with the existing approaches in hybrid and KB source prediction and also QA tasks has shown a significant reduction in response time and increased accuracy.

SCHEMA AND CONTENT AWARE CLASSIFICATION FOR PREDICTING THE SOURCES CONTAINING AN ANSWER OVER CORPUS AND KNOWLEDGE GRAPHS

SOMAYEH ASADIFAR*, MOHSEN KAHANI†, SAEIDEH SHEKARPOUR‡

5
6
7

*,†*Engineering Department, Ferdowsi university of Mashhad*
Mashhad, 9188374518, Iran‡
‡*Department of Computer Science, University of Dayton,*
Dayton, Ohio

10
11

somayyeh.asadifar@stu-mail.um.ac.ir
†*Kahani@um.ac.ir*
‡*sshekarpour1@udayton.edu*

14
15
16
17
18
19
20
21
22
23

Today, several attempts to manage question answering (QA) have been made in three separate areas: 1) Knowledge-based (KB), 2) Text-based and 3) Hybrid, which takes advantage of both prior areas in extracting the response. On the other hand, in question answering on a large number of sources, source prediction to ensure scalability is very important. In this paper, a method for source prediction is presented in hybrid QA, involving several KB sources and a text source. In a few hybrid methods for source selection, including only one KB source in addition to the textual source, prioritization or heuristics have been used that have not been evaluated so far. Most methods available in source selection services are based on general metadata or triple instances. These methods are not suitable due to the unstructured source in hybrid QA. In this research, we need data details to predict the source. In addition, unlike KB federated methods that are based on triple instances, we use the behind idea of mediated schema to ensure data integration and scalability. Results from evaluations that consider word, triple, and question level information, show that the proposed approach performs well against a few benchmarks. In addition, the comparison of the proposed method with the existing approaches in hybrid and KB source prediction and also QA tasks has shown a significant reduction in response time and increased accuracy.

Keywords: Question Answering; Hybrid Search; Data Integration; Knowledge-based and Corpus; Source Prediction.

1. Introduction

28
29
30
31
32
33
34
35
36

An ideal question answering (QA) system is a system that can display a response to the user within a reasonable time. The exploiting answer poses further challenges if it requires a hybrid approach [5], meaning not all questions are answered by structured or unstructured sources alone. Considering example question Q_1 in Table 1, the first sub-question, sq_1, is mapped to the query (?x, dbo:musicComposer, ?uri) which searches encyclopedic information for the music composer of the film. Thus, the answer to this query is extractable from the LOD cloud (DBPedia). On the other hand, the second sub-question, sq_2, specifically its phrase, "the early life of", is not encyclopedic information and so cannot be extracted from the LOD cloud. As a result, the sub-question, sq_2, should be searched using unstructured sources. Thus, to meet requirements in real situations, QA interfaces must aggregate information from both KB dataset (knowledge graph (KG)) and the corpus. Recently, much research has been conducted in this area, but all of these studies have utilized only one predetermined dataset to extract the answer [1]–[3]. Therefore, in this category of tasks, the prediction and selection of the source is not necessary task before starting to query.

39
40
41
42
43

The current work conducts a hybrid search of a large number of structured datasets and unstructured corpus. The proposed method, hybrid SP (hybrid Source Prediction), is the first solution to predict the data source in hybrid QA, involving several KG sources and a text source. hybrid SP examines the question from two views of question words and question triple. The first view considers the question as a sequence of words. Question level information is used to determine the answer source in two categories: structured or unstructured, while triple level information is used to determine the structured source containing the answer. Only a handful of hybrid QA systems refer to the choice of source from the two sources of knowledge graph and textual source. These systems have used prioritization [2]

†Corresponding author

45 between these two sources or simple heuristics [4] to select the relevant source. But so far, no evaluation has been
46 provided on execution time and accuracy for source selection. In addition, in these systems, resources are limited and
47 pre-determined.

48 Among the sources of the knowledge graph (KG), source selection methods have been proposed, regardless of its
49 application, in the topic of data integration that is represented in three main approaches: materialization [5], [6],
50 mediation [7], [8], and virtual [9], [10]. The mediation approach, among these approaches, creates a balance between
51 the cost of storage and time response. This method establishes an interface based on a mediated schema [7], [8], [11]
52 for dataset mapping.

53 On the other hand, discovering the relevant dataset on the topic of data integration falls into three categories: metadata-
54 based [12], [13], content-based [14], and link-based [15]. Content-based data integration systems are also divided into
55 instance-based [16], schema-based [17], [18] and hybrid categories [14]. In the proposed method, due to the presence
56 of unstructured source, general metadata is not sufficient to identify the source. According to the example question Q_2
57 in the Table 1, the information about "Jean-Paul Sartre" is present in the structured semantic dataset "DBpedia", while
58 the property with the concept of "refused" is not available and must be extracted from the unstructured data source.
59 Therefore, the answer to the sub-question sq_1 should be extracted from the unstructured data source.

60 Since most datasets have a very large number of triples, searching at the instance level takes a lot of time. Therefore,
61 based on the fact that in any structured dataset , the number of triples at schema level is always much less than the
62 number of triples at the instance level, so the method proposed in this paper for predicting the source is based on the
63 schema of the datasets. Among the existing methods, the most similar to the hybridSP used to select the source in the
64 QA process is LODsyndesis [14]. The method of this system is to focus on entity-based triples and consider the data
65 source structure as separate indices of class, property, and entity, independently in each data source. These two factors
66 have led to slowing down as well as low accuracy for complex questions, especially multi-hop questions in
67 LODsyndesis.

68 According to the above, in the proposed method, with the aim of increasing the speed of predicting the relevant source,
69 the data integration method based on mediated schema is used. Taking the idea from [8] about the usage of multiple
70 schemas for data integration, the current work builds the two levels data integration framework for dataset prediction.
71 Schema matching against schemas is perform using hMatcher [11].

72 On the other hand, with the increase in the size and number of LOD datasets, they became a field of interest. In general,
73 these systems can be divided into two general categories: 1) semantic parsing methods [19]–[21], that lack sufficient
74 flexibility, and 2) neural network methods [22], [23], encrypting questions and sources in vector space, which have
75 led to increased accuracy and speed.

76 Similar to existing hybrid systems, KBQA systems often perform on a limited number of predetermined datasets.
77 QAnswer [16], [24] and LODQA [21] on the other hand, are KBQA systems implemented on a substantial portion of
78 LOD. Lack of proper source selection method has been introduced as a challenge in QAnswer [16], [24]. LODQA
79 [21] uses the LODsyndesis system to select the source containing the question entity. LODQA is a semantic parsing
80 method based on Bag-of-Words. In this research, it has been used as a basic system. To increase the efficiency, the
81 BERT-based QA [23] system has been used along with the proposed system to select the appropriate source. BERT
82 [25] is a powerful method that has emerged as the latest version of neural network methods.

83 The effect of adding the proposed method for source selection along with QAnswer during response extraction is also
84 investigated.

85 Due to the KG datasets integration based on the middle schema in the proposed method, it is necessary to extract the
86 question schema to search for the answer source. Recently, the issue of sequence tagging [26] has been considered for

87 various domains. The present research for question schema generation has been influenced by the sequence tagging
88 work [27]. That study merges two well-known sequence tagging approaches, LSTM and CRF, in a type extraction
89 module. The current study extracts the type of subject and object of the given question (question schema) by
90 referencing the author's idea in [27], but does so using the BI-LSTM-CRF approach presented in [28] over BERT
91 vectors to increase the performance [29].

92   The present paper's main contribution is to provide a solution to predict the source (s) containing the response from
93 a number of structured data sources and an unstructured data source. Other contributions are as follows:

94 • Using question words information and question structure to determine the source containing the answer from the
95   KB sources and a corpus.
96 • Using the idea of data integration based on creating a mediated schema to predict the data source (s).
97 • Extraction of the question schema by the usage of sequence tagging according to the CRF and LSTM approach.
98 • Investigation of linguistic properties of questions with answers in the text by clustering BERT embedding vectors.
99 • Comparison of the proposed method based on different settings on several benchmarks.
100 • Comparison in the proposed data integration method with the state-of-the-art method for selecting the appropriate
101   KB source on simple and multi-hop questions.
102 • Comparison of accuracy and speed of existing KBQA systems implemented on a large number of KB sources
103   using the proposed data integration method.
104
105 The rest of the current paper is organized as follows. Section 2 reviews the related works, while Section 3 introduces
106 the necessary preliminaries. The proposed approach is presented in Section 4. Section 5 provides the empirical results
107 and closes with concluding remarks and future works.

108 **2. Related Works**

109 The literature related to the approach of the current research is spread over two areas: (i) hybrid question answering
110 systems and (ii) data integration, question answering and sequence tagging in LOD. The following provides a brief
111 overview of the state-of-the-art of these areas.

112 **2.1. *Hybrid Question Answering Systems***

113 Hybrid QA approaches consist of multiple structured and unstructured sources to search for an answer. Recently, some
114 hybrid QA works have developed a collaborative strategy. In text source annotated with KB entities, Park et al.
115 employed SPARQL queries if the text strategy is not successful [2]. Other works leverage text to enrich the vocabulary
116 of the structured knowledge [30], [31] or leverage KB facts to enrich the unstructured source [30], [32]. Recent systems
117 operate on text and KB continuously [3], [33], such as turn the data from various types of sources into a special global
118 representation and perform an expanded approach using memory networks [1], [34]. The major weaknesses of this
119 work which lack the rich relationship framework between information and text excerpts are solved by GRAFT-Nets
120 [3] and its next system PullNet [33].

121 The critical point in all hybrid QA systems is that all are based on utilizing one structured source and one unstructured
122 source for extracting the desired response. To extract answers, existing hybrid query systems do not provide a process
123 for selecting the appropriate resource from multiple existing (structured or unstructured) sources.

124 **2.2. *Data Integration, Question Answering and Sequence Tagging in LOD***

125 **Data Integration:** Regarding query processing in LOD, another critical issue is data integration. LOD provides a set
126 of datasets to make finding a response and data connectivity easier. Therefore, the need for data integration is
127 absolutely essential, because datasets are constructed independently and according to different principles.

128   There are three main approaches for data integration [35]. The first approach is materialization [36], in which datasets
129   are physically collected in the same place, thus increasing the response speed. However, many costs are incurred to
130   incorporate the data as well as to update possible variations. The second approach is mediation [7], [8], in which one
131   or more schema or ontology is regarded as the interface between the user and the distributed datasets. A query is first
132   executed on the interface and then on the selected datasets. In mediation, the response rate is lower than that of the
133   materialization method, but there is no cost for collecting data. The third approach is the virtual method [37], which
134   is federated query processing featuring two categories of SPARQL endpoints (e.g., LODatio [37]) and LDF clients
135   (linked data fragment) [9].

136   Unlike materialization, the virtual approach does not need to pay for the datasets' warehouse storage costs nor does it
137   need to use schemas and ontologies for the interface. However, the virtual response rate is far less than that of
138   materialization and mediation [9]. Therefore, the mediation method generates a balance between cost and execution
139   time. Thus the present research selects the mediation method.

140   Data integration approaches, on the other hand, fall into three general categories: metadata-based [12], [13], content-
141   based [14], and link-based [15], with the goal of selecting the appropriate source [35]. Another hot field of research
142   at the present time that is close to the field of data integration is called graph summarization. The purpose of this field
143   is to calculate a meaningful and concise summary of key information in a graph [17], [18]. This goal is very similar
144   to the motivation of this research, with the difference that our goal is to gather data on summaries obtained from
145   several graphs.

146   In existing data integration approaches, metadata includes descriptions of data sources, statistics, and general
147   information. Because the text source is present next to the KB datasets, metadata-based methods such as [12], [13]
148   are not applicable in the current research. The reason for this is that the answer with the same subject may not be
149   present in the KG but can be found in the text. Content-based systems are also divided into instance-based [16],
150   schema-based [17], [18] and hybrid categories [14].

151   The triples of each structured dataset in LOD are added at the instance level based on a specific schema. Since most
152   datasets have a very large number of triples, searching at the instance level takes a lot of time. Therefore, based on the
153   fact that in any structured dataset, the number of triples at schema level is always much less than the number of triples
154   at the instance level, so the method proposed in this paper for predicting the source is based on the schema of the
155   datasets.

156   Among the existing methods, the most similar to the hybridSP used to select the source in the QA process is
157   LODsyndesis [14]. This system is entity-based and extracts the sources containing the entity provided within the LOD.
158   This method is slow because of the need to search for instance level triples to determine the source containing the
159   response. In addition, for each source, it focuses on its triples, entities, properties and classes separately, so it performs
160   well only on simple questions and has poor performance on multi-hop questions that are highly regarded in research
161   today. In the present study, LODsyndesis has been used as a baseline for comparison with the proposed system.

162   On the other hand, previous research has shown that ontology [8] or schema-based [7], [11] methods handle the
163   challenges of the mediation data integration approaches with flexibility and simplicity [7]. Because ontology includes
164   the classification of classes, their relations and instances, the current work chooses schemas for the simplicity of data
165   integration in dataset selection.

166   Since authors in [8] use multiple ontologies for data integration, the present study takes that idea to manage multiple
167   datasets. The current work addresses multiple schemas with hMmatcher's schema matching [11] approach, which has
168   a higher accuracy than other methods.

169   **Knowledge-Based Question Answering**: With the increase in the size and number of datasets in the LOD, research
170   on improving the efficiency of systems in information processing as well as increasing the speed and accuracy of
171   response has expanded. With this explanation, the existing systems fall into two general categories. In the first
172   category, the systems are based on the semantic analysis of the given question, which do not have enough flexibility
173   [19]–[21]. The second category increases the efficiency of the system in finding accurate answers by encrypting
174   questions and KG in vector space [22], [23]. In the second category, the powerful BERT method is introduced as a
175   recent representation, which is used with a transformer structure and in a fine-tuned manner. The strength of BERT
176   has been evaluated in research over traditional methods [38]. Therefore, using the rich semantic properties of BERT
177   word embedding will increase system performance.

178   Although much research has been conducted in the context of QA in LOD [39], [40], there is still a very small number
179   of systems which execute queries over a large number of datasets. QAnswer [16], [24] and LODQA[21] are systems,
180   which recently implemented on a significant part of LOD. In QAnswer [16], [24], scalability is not addressed and the
181   challenge is to choose the correct datasets. The LODQA [21] system uses the LODsyndesis [14] data integration
182   system to select the appropriate source. LODsyndesis is a content-based method (hybrid of schema and instance levels)
183   to select the appropriate source. As stated in the data integration related works section, LODsyndesis [14] has been
184   used in this research as a basic system in the evaluation section. In this research, the prediction of the source containing
185   the answer is considered. Therefore, the use of the proposed system in the response extraction process is also
186   considered. The LODQA [21] system, which uses LODsyndesis [14], is also considered as one of the basic systems
187   for performing evaluations. LODQA [21] is a semantic analysis method, so to increase the efficiency of this research,
188   a powerful BERT-based system [23] is used along with the proposed source method in the evaluation section.

189   **Sequence Tagging:** Lately, sequence tagging problems have attracted much attention. Recent work has featured long-
190   term dependency learning employing Long Short-Term Memory networks (LSTM). Effective use can be made of past
191   and future characteristics for individual positions (via backward and forward states) in the bidirectional LSTM network
192   [26]. Another way to utilize tag information and generate greater tagging accuracy is applying the Conditional Random
193   Field (CRF) for the usage of sentence level data. Inspired by [27], the current work uses CRF with the LSTM method
194   to identify the type of subject and the object of the given question. The present study for question schema generation,
195   adopts the method introduced by [28], with the BI-LSTM-CRF approach.

196   The current paper intends to examine the gap between these two considered related work areas. This is achieved by
197   introducing a continuous and simultaneous prediction method for selecting sources that contain the answer among
198   multiple KB datasets and textual data sources.

199   **3. Preliminaries**

200   This section presents the necessary preliminaries. To promote the importance of the KB scheme in the current
201   approach, the general definition of the database schema is first formally presented, followed by the definition of the
202   KB dataset and its schema. Throughout the present paper, the term dataset refers to KB databases and the term source
203   signifies both the KB and corpus databases. In addition, the knowledge graph (KG) and the knowledge-based (KB)
204   dataset are used interchangeably.

205   **3.1. *Database Schema***

206   The term database schema refers to the structure defined in a formal language which organizes the database and
207   maintains its integrity. In other words, data is arranged within a schema and, overall, the schema is the main body of
208   the database that determines how the database is to be constructed.

209   **3.2. *RDF***

210   An RDF (Resource Description Framework) is a metadata model specified in the W3C RDF recommendation§ for
211   building an infrastructure of machine-readable semantics for data on the web. In the RDF model, the universe to be
212   modeled is a set of resources, that is, essentially anything that has a universal resource identifier, URI. The RDF graph,
213   G, is a set of triples with a structure (i.e., subject, predicate, object) abbreviated as $(s,p,o)^{**} \in (U \cup B) \times (U) \times (U \cup$
214    $B \cup L)$, in which U represents URI references, L stands for literals, and $B \subset (U \cup L)$ is for blank nodes††. The
215   predicate, p , is a relation between s and o. Each triple is represented by the labeled graph, $s \xrightarrow{p} o$. The RDF database
216   is named Knowledge Graph (KG).

217   **3.3. *RDF Schema (KG Schema)***

218   RDF Schema S is abbreviated as RDFS, RDF(S), RDF-S or RDF/S, associated with the RDF graph, G, which is the
219   built-in vocabulary, $voc(G) \subset U \cup L$, has an inheritance of classes, C, and certain properties, P, and provides basic
220   elements for the ontology definition. Properties, denoted as a predicate, describe a relation between subject s and
221   object o resources. S is a set of triples (sc,p,oc), in which sc indicates the class of the subject and oc represents the
222   class of the object for the triple, whose predicate is property p. The proposed method involves two types of
223   connections: Property-based connections and ontology-based connections. The first involves connections that have a
224   property between two different classes (e.g., (dbo:PopulatedPlace, dbo:officialLanguage, dbo:Language)). The latter
225   involves the connections that exist between the various elements in an ontology. For example, the relationships
226   "subclass-of", "part-of" can be mentioned (See Fig. 1).

227   Fig. 2 illustrates three triples: (i) (dbr:Italy, dbo:officialLanguage, dbr:Italian_language), (ii) (dbr:Italy, rdf:type,
228   dbo:PopulatedPlace), and (iii) (dbr:Italian_language, rdf:type, dbo:Language). In the second triple,
229   "dbo:PopulatedPlace" expresses the class of the resource, dbr:Italy, and, in the third triple, dbo:Language is the class
230   of the resource, dbr:Italian_language. The related schema for the first triple in the RDF graph, G, is represented in
231   Schema S, while the property, dbo:officialLanguage, describes a relation between the subject resource associated with
232   the class of dbo:PopulatedPlace and the object resource associated with the class of dbo:Language.

233   The following is a formal definition of the question schema concept in the current paper.

234   **3.4. *Question Schema***

235   The schema of the triple (s,p,o) is defined as (sc,p,oc), in which sc represents the class of subject s and oc stands for
236   the class of object o by considering an underlying ontology. Since the topology and vocabulary of background
237   knowledge graphs are heterogeneously declared in various domains, the direction of the extracted schema depends on
238   the background schema or ontology. Therefore, in order to have a generic model, both directions of an extracted
239   schema, (sc,p,oc) and (oc,p,sc), are considered.

240   **3.5. *Baseline systems***

241   Since in this study the proposed data integration method, hybridSP, is compared with LODsyndesis [14], the indices
242   generated in this method are shown in Fig. 3. The main purpose of LODsyndesis is to identify data sources containing
243   the desired entity. Thus, entity, property, constant, class, and entity-based triple indices are generated.

---

§ https://www.w3.org/TR/rdf-concepts/#section-Graph-syntax.

** This definition is inspired by RDF triple.

†† This feature is useful when representing complex attributes, such as address.

244   Since in this study the aim is to discover a suitable data source to extract the answer to the given question, so we will
245   examine its application in the question answering system. Recently, the LODQA question answering system [21] uses
246   the LODsyndesis [14] content-based data integration method to identify appropriate datasets. Figure 4 shows the
247   architecture of the LODQA system.

248   The proposed method for extracting the question schema is based on BERT and through LSTM. To compare hybridSP
249   in the QA process, a KBQA based on BERT is used [23]. This system is called bertQA for short in this research.
250   Figure 5 shows the bertQA system architecture. The two boxes on the right and left (with dashed line) run in the
251   offline section and prepare the models needed for the online section.

4.   Approach

253   The proposed approach, hybridSP, addresses the challenge of source prediction in the context of several KB datasets
254   and a corpus. The purpose of source prediction in current research is to find the best source containing the answer.
255   Therefore, the hybridSP in the question answering process is presented in Figure 6. The box (with dashed line) on the
256   left of Figure 6, which includes offline and online parts, shows hybridSP. If sources from the KG are predicted, the
257   proposed method uses the bertQA system [23] to search for answers. If the text source is identified as the source
258   containing the answer, the Lucene is used to extract the answer.

259   Details of the hybridSP are described below.

4.1.   *Offline part*

261   The offline section, MUGschema index, provides the required platform for the online part. The main task in the offline
262   part is to store the schema of the KB datasets so that in the online part, the process of predicting the response source
263   is done at high speed. Fig. 7 shows the tasks required in the offline part, which include the following steps.

264   4.1.1.   ***Generating the mediated upper global schema:***

265   Fig. 8 illustrated the structure of creating an intermediate layer in the proposed method, taking ideas from data
266   integration methods [8]. The production of the mediate layer includes the following steps: (1) categorizing the datasets
267   according to the domain under study, (2) collecting the local schema of the datasets in each domain, (3) integrating
268   the local schemas of each domain, and (4) integrating the global domain schemas and creating the mediated global
269   upper schema (It is the upper level in Fig. 8). Part of the mediated upper global schema generated between the three
270   biomedical datasets[‡‡], including Drugbank, Sider, and Diseasome, is shown in Figure 9.

271   4.1.2.   *Generating an Embedded-Mediated Upper Global Schema Index:*

272   This step consists of the following three minor steps. The first step is Mediated upper global schema expansion, which
273   in this section for each of existed triple in the form of (object, predicate, subject) (e.g., T1 in Fig. 10) that have an
274   object neighbor triple (e.g., T_2 in Fig. 10) and a subject neighbor triple (e.g., T_3 in Fig. 10), three types of sequences
275    $T_2 + T_1$, $T_1 + T_3$ and $T_2 + T_1 + T_3$ are produced. The purpose of expanding the generated schema in last step, is to increase
276   the efficiency of the system for multi-hop questions, because to find the correct answer to such questions in KBs, it is
277   necessary to infer several neighboring triples. The Table 2 contains a 1-hop question and two types of multi-hop
278   questions. The first question, Q_1, is a 1-hop and can only be answered with a triple. The second row shows a question,
279   Q_2, belonging to the MetaQA [41] database, which can be answered by inferring several neighboring triples from one
280   KB dataset. The third row, Q_3, is a question from the biomedical QALD [42] dataset that to answer the question,
281   triples from two different KB datasets must be considered. Both questions are 3-hop and require inference on three

---

[‡‡] http://qald.aksw.org/index.php?x=task2&q=4

282 triples to arrive at the correct answer. The expansion of each triple with its neighbors will increase the accuracy of the
283 system in selecting the appropriate sources containing the answers to the multi-hop questions.

284 Since the generated schema is the result of the integration of several schema datasets, the neighboring triples that
285 merge with each other for expansion, may belong to different datasets that must be considered in the indexing step.
286 According to the Fig. 10, it can be seen that T_1 and T_3 triples belong to Disasome and Drugbank datasets, respectively,
287 so the resulting sequence T_{1+} T_3 belongs to both datasets.

288 After expanding triples, the embedded vector of the generated sequences is created in second step and stored in the
289 index file, MUGschema index, along with the identifier of the related datasets in third step. Table 3 shows part of the
290 index stored for the information provided in Figure 10.

291 **4.2. *Online part***

292 The present study employs two main modules: one, in a triple level, to extract features for comparison with LOD
293 datasets (with an schema in mind) and one, in a question and word level (with content in mind) to extract features for
294 comparison with the corpus.

295 These two modules are combined with other modules, Similarity Calculation, Binary Classification, and Multi-
296 Labeled Classification, to predict the source(s) of the answer defined in the following discussion.

297 To select the source(s) of the answer, the current work proposes two variations containing two main modules, triple
298 level and question level, in a serial or parallel combination with each other and other modules. In the first variation,
299 the triple level module works in parallel with the question level module. However, in the second variation, the triple
300 level module works in a serial combination in addition to a parallel combination and its output is concatenated in the
301 question level module. Figs. 11 and 12 present the graphical representation of the two proposed variations.

302 4.2.1. *Triple Level Module*

303 This section introduces a question schema generation model for dataset prediction. The question schema, e.g., (person,
304 parent, person), carries very useful information for comparison with the dataset schema. The inspiration of this model
305 is the work by [28], which proposes using the LSTM network method to extract the answer type.

306 **Decomposing Question and Generating Question Triple:** The first step in the Triple Level module is question
307 decomposition and converting each sub-question into triple form (subject, relation, object) for use in the next step,
308 Generating Question Schema .

309 For question decomposition and triple generation, the existing syntax and pattern-based work [20] are employed. After
310 generating the triple of the given question, the current work's objective is to construct the schema of the triple.
311 Sequence tagging is used to solve the question schema construction problem.

312 The key idea is to view each predicate/entity as a low-dimensional vector so that the relation information in KB may
313 be retained. The proposed solution, therefore, integrates the low-dimensional representation in the first layer in order
314 to transform a term from a character series into a dense vector sequence. Then, bi-LSTM [28] and CRF are used for
315 producing the query schema to predict the datasets containing the answer. With such combining layers, past and future
316 tags can be efficiently used to predict the class of the object and subject of the question.

317 **Generating Question Schema:** In order to reach a question schema, a confidence score for each possible KB type of
318 the generated object, predicate, and subject must be computed via the bidirectional LSTM and CRF approaches [28].
319 Figure 13 illustrates the proposed model for question schema construction.

320   It is supposed that the set of all class types in the underlying KB(s) is $T = (y_1, y_2, ..., y_k)$. The given mentions are the
321   question words, $q = \{x_1, x_2, ..., x_n\}$, and the generated subject, predicate, and object from the last step are s, p, and o,
322   respectively. For the given mention, m , with the starting i and ending j indices, the proposed approach, inspired by
323   [28], employs an average pooling layer to generate the type of the mention. In the proposed model of the schema
324   construction in Figure 13, the hidden layer in the LSTM step computes the score matrix, $S_{3 \times k}$, in which three rows
325   are for the mentions of subject, predicate and object. $s_{i,k} \in R$ denotes the likelihood of y_k as the correct type of m in q.

326   At first, LSTM is learned from left to right and right to left over embedding vectors of tokens, while the forward and
327   backward outputs, $\overleftarrow{h}$ and $\overrightarrow{h}$, are computed. For all output nodes, $r \in [i,j]$, the backward v_r, forward v_r, backward v_m,
328   and forward v_m for mention m and the final representation of v_m (denoted for concatenation by $\oplus$) are obtained by
329   the following equations:

$$\overleftarrow{v}_r = \frac{1}{r}\Sigma_{i=1}^{r} \overleftarrow{h}_i \tag{1}$$

330

$$\overrightarrow{v}_r = \frac{1}{n-r+1}\Sigma_{i=r}^{n} \overrightarrow{h}_i \tag{2}$$

331

$$\overleftarrow{v}_m = \frac{1}{j-i+1}\Sigma_{r=i}^{j} \overleftarrow{v}_r \tag{3}$$

332

$$\overrightarrow{v}_m = \frac{1}{j-i+1}\Sigma_{r=i}^{j} \overrightarrow{v}_r \tag{4}$$

333

$$v_m = \overleftarrow{v}_m \oplus \overrightarrow{v}_m \tag{5}$$

334

335   The obtained score matrix, $S_{3 \times k}$, including the output nodes for the three mentions, v_s, v_p and v_o, in sentence x enters
336   the CRF layer to receive the probability distribution of class types. CRF is based on the graph, $G = (V,E)$, in which, V
337   is a set of random variables, Y_i, E is an edge between Y_{i-1} and Y_i, f_k is a binary feature function, k is the number of
338   defined edge variables, and k' is the number of vertices and the two factors considered for vertices V, F_V and edges E,
339   F_E (Eqs. (6) and (7)). S(x,y) is the sum of the two factors (Eq. (8)) (x is the sequence of tokens v_s, v_p , and v_o, and y
340   is the sequence of tags). The conditional probability $p(y|x)$, is obtained by Eq. (9). The goal is to maximize the
341   probability of the distribution function results in $y^* = \text{argmax}_{y' \in y} (s(x,y'))$.

$$F_V = \Sigma_{i=1}^{3} \lambda'_k f'_k(y_i, x) \tag{6}$$

342

$$F_E = \Sigma_{i=1}^{3} \lambda'_k f'_k(y_{i-1}, y_i, x) \tag{7}$$

343

$$s(x,y) = F_V + F_E \tag{8}$$

344

$$p(y|x) = \frac{e^{s(x,y)}}{\Sigma_{y' \in y} e^{s(x,y')}} \tag{9}$$

345

346   ### 4.2.2. *Question Level Module*

347   For all words, w_i, in the given question, q_i, a vector of features, V_{wi}, is generated. Two approaches are proposed:

348   (1) Language Feature vector (LF): This vector features properties about parts of speech tagging (POS), dependencies,
349   and the semantic role labeling of each word, w_i.

350   (2) Embedding Feature vector (EF): These word embedding vectors are pre-trained using BERT on Wikipedia pages
351   and BookCorpus. For each word, a 768-dimensional vector is obtained. The question vector is computed with the
352   average function over the word vectors.

353   ### 4.2.3. *Similarity Calculation Module (SC)*

354   For each underlying KB, G, the schema set, $S = \{s_1, s_2, ..., s_n\}$, is prepared based on the definition in Section 3.4. The
355   average embedding vectors for the schema set $S_e = \{s_{e1}, s_{e2}, ..., s_{en}\}$, and the given question schema, v_{qs}, are calculated
356   with the average function, AVG in Eq. (10). A maximum reverse distance, D to $sim(v_{qs}, G)$, is assigned in the Eq. (11).

$$v_{qs} = AVG(v_s, v_p, v_o) \tag{10}$$

$$sim(v_{qs}, G) = max_{se \in S_e} e^{\frac{1}{D(v_{qs}, se)}} \tag{11}$$

359   ### 4.2.4. *Binary Classification Module (BC)*

360   The binary classification module classifies the vector of the given question based on its textual features in the RDF or
361   textual classes.

362   ### 4.2.5. *Multi-label Classification Module (MLC)*

363   In previous modules based on the applied model, two categories of data are extracted: (1) calculated similarity scores,
364   $SS = \{sim_{G1}, sim_{G2}, ..., sim_{Gm}\}$, for m underlying KBs, which determine the similarity score of the question schema to
365   underlying KBs and (2) the class of the source of the answer (structured (RDF) or unstructured (textual)). In this
366   module, the aim is to combine textual and underlying KB features to predict the sources(s) containing the answer. The
367   current research assumes that its underlying sources are one or more knowledge graph (s) and one corpus. In the
368   context of a hybrid search, the answer of the given question is extracted using one or more source(s), considered as
369   the question class labels. As more than one class label for each given question is considered, a multi-labeled
370   classification (MLC) is a good fit as the solution for the source prediction problem allowing multiple predictions for
371   several classes at the same time.

372   ## 5. Experiments

373   This section describes the benchmarks, data preparation, and experimental results. Performance is evaluated in terms
374   of the three main tasks contributed by the proposed approach: (i) question schema extraction (triple level), (ii) binary
375   or question level classification, and (iii) source prediction (a combination of the triple level and question level). In his
376   study, the purpose of source selection is to find suitable sources for extracting answers. Therefore, the impact of the
377   presence of source prediction method in the QA system is finally evaluated. All experiments are conducted on a
378   machine with Nvidia Tesla v100 and p100 GPUs.

379   The accuracy is reported as the evaluation metric. Accuracy indicates the ratio of the number of correctly identified
380   samples to the total number of samples. For each question, q_i, let T be the true set of labels and S be the predicted set
381   of labels. Accuracy is measured by the Hamming score which symmetrically measures how close T is to S (Eq. (12)).

$$\textit{Accuracy } (q_i) = |S \cap T| / |S \cup T| \tag{12}$$

383   **5.1.   *Benchmarks***

384   To assess the tasks in the literature's related works, two types of benchmarks are employed: (i) benchmarks used for
385   question answering systems based on KBs, such as simpleQuestions [43], webQuestions [44], webQuestionsSP [33],
386   ComplexQuestions [45], ComplexWebQuestions [46], WikiMovies [34], MetaQA [41], LC-QUAD, and LC-QUAD
387   2.0 [47], multi-hop Biomedical datasets in QALD challenge[§§] [42] and (ii) benchmarks used for hybrid question
388   answering systems, such as SPADES [48] and hybrid dataset in QALD challenge [49].

389   For its evaluation of schema extraction, question level classification, and source prediction, the current work employs
390   the ComplexWebQuestions, which is constructed over webQuestions with additional constraints and SPARQL logical
391   forms. webQuestionsSP and ComplexQuestions are also constructed over webQuestions and WikiMovies has the
392   same properties as webQuestions. Thus, the present study does not consider these benchmarks. LC-QUAD is a
393   benchmark in the current study and used in QAnswer (One of the basic QA systems in present research) as well. The
394   three tasks are evaluated against hybrid QALD benchmark. In the section of source prediction evaluation as well as
395   answer extraction, three categories of simple questions containing SimpleQuestions [43], multi-hop questions
396   containing biomedical QALD [42] and MetaQA [41] and hybrid questions containing hybrid QALD [49] and SPADE
397   [48] benchmarks are used.

398   Table 4 lists the general characteristics of the benchmarks used.

399   **5.2.   *Data Preparation***

400   Since the current study's approach is based on schema dataset integration, the schema of the underlying knowledge
401   graphs (s) must be determined. As seen in Table 4, the benchmarks are related to DBpedia, Freebase, Wikidata, and
402   CluWeb, Sider, Diseasome, Drugbank and WikiMovies. Additionally, for the response time evaluation and
403   comparison with QAnswer, the questions are executed over DBpedia, wikidata, Freebase, MusicBrainz, and DBLP
404   [16], [24]. As a result, the RDF data of underlying KBs are collected in the form of triples (s,p,o) and then the type of
405   subject s and object o are retrieved to prepare the schema of the KGs. The state-of-the-art approach, hMmatcher [11],
406   performs the schema matching against schema.

407   **5.3.   *Training Data***

408   Since the considered databases only provide question-answer pairs, it is necessary to figure out the type information
409   to be used as training data. Simulated types are obtained as follows: for each relation r, ($?,r,o$) and ($s,r,?$) are queried
410   based on the underlying KGs. The subject and object types are selected as related types. In this research, the PyTorch[***]
411   implementation of BERT from Huggingface has been used to encode the questions and POS tags. In the pre-training,
412   the bert-base-uncased model [25] is used. This model includes 12 transfer blocks, hidden layer with size 768, 12
413   attention layers and Adam optimizer with 110M parameter. After the vectors are obtained by BERT, for fine-tunning
414   step a linear layer for binary classification and several layers for multi-label classification stacked on top of the BERT
415   output are applied. In multi-label classification, the output of the model will be the probability of all classes. For fine
416   tuning, a learning rate of 3e-5, a batch size of 128 and 4 epochs is used.

417   **5.4.   *Question Schema Extraction***

418   For the first task of question schema extraction, the current study utilizes 768-dimensional word and parts of speech
419   (POS) embeddings, which are pre-trained using BERT on Wikipedia pages and BookCorpus. The size of the LSTM
420   hidden layer is set at 768 in a forward and backward direction. A 768-dimensional vector representation is learned for
421   each class of subject and object. RMSProp [50] is utilized with a learning rate of 0.005. Because the dropout method

---

[§§] http://qald.aksw.org
[***] https://github.com/huggingface/pytorch-transformers

422 is an efficient way of mitigating over fitting, the present research applies it to the embedding, LSTM, and CRF layer
423 outputs. The dropout rate is 0.2 and the hyper-parameters are selected on training data based on cross-validation. The
424 networks are implemented on the common Theano neural network toolkit [51].

425 For each given question, the evaluation relies on two annotators to judge the generated schema. If the output of the
426 two annotators is correctly labeled, then this is considered as the correct output. The agreement rate (Cohen's kappa)
427 is 0.80%.

428 Figure 14 illustrates the comparison of the accuracy metric among several benchmarks against different applied
429 approaches. As seen, ComplexWebQuestion and LC-QULD have a higher accuracy because their questions are
430 selected from KG (s). The schema of such questions which corresponds to the schema of underlying KGs are
431 extractable. Hybrid benchmarks (hybrid QALD and SPADES) include questions that can only be answered from the
432 text. The schema of these questions is not available in the KG. Furthermore, the use of CRF, word embedding, and
433 POS embedding leads to improved performance.

434 **5.5. *Question Level Classification***

435 The second case is binary or question level classification. The only system that refers to the choice of source between
436 a KG and a text source is the HAWK system [4]. In this system, noun phrases are searched from the text and the rest
437 from the KG which no evaluation has been provided for it. In this section, first, we will evaluate the binary class with
438 two tags, text and RDF, on different datasets and based on different settings. Then we compare the hybridSP with the
439 HAWK method.

440 The constructed BERT vector of the given question is classified based on a model fine-tunning. This model employs
441 a binary classifier layer, to classify the question in either two classes of, i.e., KG or corpus class. Figure 15 provides
442 the accuracy for classifying questions. Results are calculated by applying the first variation (FV) (Fig. 11) or the
443 second variation (SV) (Fig. 12) of the proposed model and embedding feature vectors (EF) or language feature vectors
444 (LF) and the BERT classifier or SVM classifier. Scikit-learn 0.19.2 is used for SVM and PyTorch for BERT classifier.
445 As seen in Figure 15, the large benchmarks show a better performance than those by the smaller benchmarks. In
446 addition, between the two large benchmarks and the two small ones, the benchmarks constructing questions over KGs
447 show a better performance than the hybrid benchmark.

448 As shown in previous research [38], SVM is less efficient than BERT. In terms of comparing the first and second
449 model variations, one can see that the second model's performance is better. While the binary classifier in the first
450 model only works based on the question level, the second model concatenates the triple level and question level as
451 inputs to the classifier. Furthermore, the performance of embedding feature vectors surpasses that of language feature
452 vectors. A possible explanation for this is the consideration of both semantic and syntactic features when using
453 embedding space.

454 The question classification in the hybridSP and the HAWK system on the QALD benchmark had an accuracy of 0.78
455 and 0.45, respectively. These values on the large SPADE benchmark showed 0.79 and 0.56, respectively.

456 To extract the properties involved in categorizing questions in the text category, the present study runs a clustering
457 algorithm over embedding vectors of the questions. With an analysis of each cluster by a natural language expert, the
458 categories in the first column of Table 5 are recognized. These categories include *Sequence Adverbs*, *Adverbials of*
459 *Location*, *Event indicators of Time*, *Quality Adjectives*, *Adverbs of Time*, *Quantifiers*, *Adverbs of Frequency,* and
460 *Hedges*. In the second column, a number of examples are given for each category. By comparing the embedding
461 vectors of other words with the embedding vectors of given examples, the other items in these categories are obtained.

462 The study of the extracted categories and the identification of more categories through the examination of other
463 question benchmarks should be left to future research. One can refer to [52], [53] for further work in this area.

464 **5.6. *Source prediction***

465 In this section, we first compare hybridPS with the HAWK method, for predicating the source. Then we will compare
466 hybridPS with the most similar method, LODsyndesis, that selects the source on the KG sources.

467 The source prediction in the hybridSP and the HAWK system on the QALD benchmark had an accuracy of 0.82 and
468 0.45, respectively. These values on the large SPADE benchmark showed 0.85 and 0.56, respectively. Examining the
469 result of this evaluation with the previous section (Question Level Classification) shows the effect of considering the
470 schema of the question on the accuracy.

471 As mentioned in the previous sections, in predicting the appropriate source in KGs, the most similar method to the
472 hybridSP is the LODsyndesis system. This system selects the sources available in the KG by focusing on the entity
473 provided. hybridSP, in addition to KG sources, also predicts the existence of responses in the textual source. It also
474 provides solutions for multi-hop questions to increase source predicting efficiency. Therefore, in addition to evaluating
475 the two systems on simple questions, multi-hop and hybrid benchmarks has also been used to compare the two
476 systems. By referring to Figure 16, it can be seen that both systems have the same accuracy in relation to simple
477 questions.

478 Unlike LODsyndesis, which is based on triples, hybridSP searches based on the schema of triples. The schema of
479 triplets has a much smaller number of triples. Thus, for all benchmarks, source prediction in LODsyndesis takes longer
480 than in hybridSP.

481 Two different types have been used to evaluate the two systems on multi-hop questions. In the first type (e.g.,
482 MetaQA), the triples related to each question are present in a common KG dataset, while in the second type (e.g.,
483 biomedical QALD), they are present in two or more different datasets.

484 As can be seen, the LODsyndesis system offers much less accuracy than the hybridSP system because it does not infer
485 multiple sequential relationships. The reduction in accuracy is greater in relation to the second type. This observation
486 is due to the fact that in the first type, the question entity and the answer are located in a common KG, while in the
487 second type, the two are often in different KGs. LODsyndesis focuses on entity-based triples in each KG, thus it will
488 be less efficient in case of such questions.

489 **5.7. *The efficiency of the hybridSP in the QA system***

490 In this research, the Question answering benchmarks has been used to evaluate the hybridSP to select the appropriate
491 source containing the answer. Therefore, in this study, we also evaluate the efficiency of the hybridSP in the QA
492 system. In this section, we first compare the accuracy and average response time of two QA systems, the former using
493 hybridSP and the latter using the LODsyndesis method (LODQA [21]) to select the appropriate source. Then we will
494 examine the result of adding hybridSP to the QAnswer system in the difference of response time.

495 As shown in Figure 16, the hybridSP system is more efficient than the LODsyndesis system in terms of accuracy and
496 prediction time. In this section, we will evaluate the role of two systems in the answer extraction process. Recently,
497 the LODQA system has used LODsyndesis to identify the source. Because LODQA uses traditional methods in
498 question representation to extract answers, this study uses powerful BERT-based question answering system [23] in
499 its evaluation.

500 As can be seen in Figure 17, the accuracy of the proposed method has been significantly improved compared to
501 LODQA due to the presence of hybridSP alongside the powerful BERT-based QA system. The remarkable thing about
502 Figure 17 is the low accuracy of the BERT-based QA system for multi-hop questions. The reason is that this system
503 works in a single relational category and does not deal with multi-hop questions.

504 We compare the response time within the QAnswer framework with and without the source prediction. In QAnswer,
505 questions are executed over DBpedia, wikidata, Freebase, MusicBrainz, and DBLP. The QAnswer system reports an
506 average response time of 1.37 seconds for the hybrid QALD benchmark when executed over DBpedia. If this is
507 executed on all underlying datasets (DBpedia, wikidata, Freebase, MusicBrainz, and DBLP), the need to run QAnswer
508 on all datasets was eliminated and only on predicted resources containing responses. With the addition of hybridSP
509 alongside QAnswer, the average response extraction time was reduced to 5.8 seconds. The average of source
510 prediction time by hybridSP on the stated datasets is 4.11 seconds. While the response time average by QAnswer on
511 the predicted sources is 1.69 seconds. The increase of about 0.3 seconds compared to the QAnswr runtime on DBpedia
512 is due to the fact that for some questions, sources in addition to DBpedia containing the answer has been identified.

513 **6. Conclusion and Future Work**

514 To extract a final answer, a question answering committee has recently been presented to hybrid approaches by
515 integrating structured (KG) and unstructured (corpus) sources. Also, many studies have set out to extract the response
516 from LOD. However, the majority of the work already conducted on datasets which are pre-determined and limited.
517 In QA on a large number of sources, it is very important to choose the relevant sources to maintain scalability. So far,
518 no method has been proposed for source selection in hybrid sources. In KG, the data integration domain chooses the
519 source among KGs. In this study, a data integration-based method with a content-based mediation approach is
520 presented. The proposed method, hybridSP, includes three main tasks of (i) generating a question schema (triple level),
521 (ii) question classification (word and question levels) and (iii) combining the results of the first and second parts in
522 identifying the source containing the answer. The hybridSP aims to identify the source containing the answer.
523 Therefore, in this research, the hybridSP for selecting the appropriate source along with the QA system as a general
524 task has also been investigated. The mentioned tasks were evaluated using several benchmarks from three categories
525 of simple, multi-hop and hybrid questions. The evaluation results showed that the proposed system of source selection
526 compared to the methods of source selection from hybrid sources, HAWK, as well as KG sources, LODsyndesis, have
527 a higher efficiency in accuracy and execution time. For the general QA task, the BERT-based QA system was used
528 along with the proposed source selection system. A comparison of the combination of the two systems found a
529 significant increase in performance compared to the LODQA system, which uses LODsyndesis. Moreover, the
530 addition of hybridSP alongside QAnswer system significantly reduces the average response time. QAnswer has the
531 ability to respond in the context of several KG datasets. Future research shall extend the scope of this work by
532 considering a larger number of KGs and textual resources over large-scale complex questions especially multi-hop
533 benchmarks. Also index clustering containing BERT embedded vectors to reduce the prediction time is considered in
534 future work.

535 **References**

536

537 [1]   E. Dimitrakis, K. Sgontzos, and Y. Tzitzikas, A survey on question answering systems over linked data and documents,
538   J. Intell. Inf. Syst., 55 (2) (2020), 233–259.
539 [2]   S. Park, S. Kwon, B. Kim, and G. G. Lee, ISOFT at QALD-5: Hybrid question answering system over linked data and
540   text data," CEUR Workshop Proc., 1391 (2015).
541 [3]   H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. W. Cohen, Open domain question answering using
542   early fusion of knowledge bases and text, Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, (2020),
543   4231–4242.
544 [4]   R. Usbeck, A. C. N. Ngomo, L. Bühmann, and C. Unger, HAWK - hybrid question answering using linked data, Lect.
545   Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 9088 (2015), 353–368.

546 [5] M. E. Papadaki, Y. Tzitzikas and N. Spyratos. Analytics over RDF Graphs, Algorithms, 1197 (2020).

547 [6] M. E. Papadaki, N. Spyratos, and Y. Tzitzikas, Towards interactive analytics over RDF graphs," Algorithms, 14 (2) (2021),
548 1–22.

549 [7] C. Farré, J. Varga, and R. Almar, GraphQL Schema Generation for Data-Intensive Web APIs, Lect. Notes Comput. Sci.
550 (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 11815 LNCS (2019), 84–194.

551

552 [1] E. Dimitrakis, K. Sgontzos, and Y. Tzitzikas, A survey on question answering systems over linked data and documents,
553 J. Intell. Inf. Syst., 55 (2) (2020), 233–259.

554 [2] S. Park, S. Kwon, B. Kim, and G. G. Lee, ISOFT at QALD-5: Hybrid question answering system over linked data and
555 text data," CEUR Workshop Proc., 1391 (2015).

556 [3] H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, and W. W. Cohen, Open domain question answering using
557 early fusion of knowledge bases and text, Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, (2020),
558 4231–4242.

559 [4] R. Usbeck, A. C. N. Ngomo, L. Bühmann, and C. Unger, HAWK - hybrid question answering using linked data, Lect.
560 Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 9088 (2015), 353–368.

561 [5] M. E. Papadaki, Y. Tzitzikas and N. Spyratos. Analytics over RDF Graphs, Algorithms, 1197 (2020).

562 [6] M. E. Papadaki, N. Spyratos, and Y. Tzitzikas, Towards interactive analytics over RDF graphs," Algorithms, 14 (2) (2021),
563 1–22.

564 [7] C. Farré, J. Varga, and R. Almar, GraphQL Schema Generation for Data-Intensive Web APIs, Lect. Notes Comput. Sci.
565 (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 11815 LNCS (2019), 84–194.

566 [8] F. Ekaputra, M. Sabou, E. Serral Asensio, E. Kiesling, and S. Biffl, Ontology-based data integration in multi-disciplinary
567 engineering environments: a review, Open J. Inf. Syst., 4 (1) (2017), 1–26, 2017.

568 [9] K. M. Endris, M. Galkin, I. Lytra, M. N. Mami, M. E. Vidal, and S. Auer, MULDER: Querying the linked data web by
569 bridging RDF molecule templates, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes
570 Bioinformatics), 10438 LNCS (2017), 3–18.

571 [10] A. A. Algosaibi, High-Performance Computing Based Approach for Improving Semantic-Based Federated Data
572 Processing, Int. J. Math. Comput. Sci., 16 (1) (2021), 287–309.

573 [11] A. Yousfi, M. H. El Yazidi, and A. Zellou, hMatcher: Matching schemas holistically, Int. J. Intell. Eng. Syst., 13 (5)
574 (2020), 490–501.

575 [12] C. B. Neto, K. Müller, M. Brümmer, D. Kontokostas, and S. Hellmann, LODVader: An Interface to LOD Visualization,
576 Analytics and DiscovERy in Real-time, Www,(2016), 163–166.

577 [13] S. Yumusak, E. Dogdu, H. Kodaz, A. Kamilaris, and P. Y. Vandenbussche, "SpEnD: Linked data SPARQL endpoints
578 discovery using search engines," IEICE Trans. Inf. Syst., E100D (4) (2017), 758–767.

579 [14] M. Mountantonakis, LODsyndesis : Global Scale Knowledge Services, Heritage, 1 (2) (2018), 335–348.

580 [15] P. Ristoski, C. Bizer, and H. Paulheim, Mining the Web of Linked Data with RapidMiner, J. Web Semant., 35 (2015),
581 142–151.

582 [16] D. Diefenbach, V. Lully, P. H. Migliatti, K. Singh, O. Qawasmeh, and P. Maret, QAnswer: A question answering prototype
583 bridging the gap between a considerable part of the LOD cloud and end-users, Web Conf. 2019 - Proc. World Wide Web
584 Conf. WWW 2019, (2019), 3507–3510, 2019.

585 [17] A. Khan, S. S. Bhowmick, and F. Bonchi, Summarizing static and dynamic big graphs, Proc. VLDB Endow., 10 (12)
586 (2017), 1981–1984.

587 [18] A. Scherp and T. Blume, Schema-level Index Models for Web Data Search, J. Data Intell., 2 (1) (2021), 47–63, 2021.

588 [19] W. Zheng, J. X. Yu, L. Zou, and H. Cheng, Question Answering Over Knowledge Graphs : Question Understanding Via
589 Template Decomposition, Proceedings of the VLDB Endowment, 11 (11) (2018), 1373–1386.

590 [20] S. Shin and K. H. Lee, Processing knowledge graph-based complex questions through question decomposition and
591 recomposition, Inf. Sci. (Ny)., 523 (2020), 234–244.

592 [21] E. Dimitrakis, K. Sgontzos, M. Mountantonakis, and Y. Tzitzikas, Enabling Efficient Question Answering over Hundreds
593 of Linked Datasets, Commun. Comput. Inf. Sci., 1197 CCIS (2020), 3–17.

594 [22] R. Wang and Z. Ling, Knowledge Base Question Answering With Attentive Pooling for Question Representation, IEEE
595 Access 7, 7 (2019), 46773–46784.

596 [23] J. Su and S. Yu, A BERT-based Approach with Relation-aware Attention for Knowledge Base Question Answering,
597 International Joint Conference on Neural Networks (IJCNN). IEEE, (2020), 1-8.

598 [24] D. Diefenbach, A. Both, K. Singh, and P. Maret, Towards a question answering system over the Semantic Web, Semant.
599 Web, 11 (3) (2020), 421–439.

600 [25] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language
601   understanding, NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. -
602   Proc. Conf., 1 (Mlm) (2019), 4171–4186.

603 [26] R. Panchendrarajan, Bidirectional LSTM-CRF for Named Entity Recognition, Proceedings of the 32nd Pacific Asia
604   Conference on Language, Information and Computation,(2018).

605 [27] K. Mrhar, O. Douimi, M. Abik, and N. C. Benabdellah, Towards a semantic integration of data from learning platforms,
606   IAES Int. J. Artif. Intell., 9 (3) (2020), 535–544.

607 [28] S. Yavuz, I. Gur, Y. Su, M. Srivatsa, and X. Yan, Improving semantic parsing via answer type inference, EMNLP 2016 -
608   Conf. Empir. Methods Nat. Lang. Process. Proc., (2016), 149–159.

609 [29] I. Elgendy, "Fine-tuned BERT Model for Multi-Label Tweets Classification. TREC. 2019, (2019), 1-7.

610 [30] W. Xiong, H. Wang, and W. Y. Wang, Progressively pretrained dense corpus index for open-domain question answering,
611   EACL 2021 - 16th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Conf., (2021), 2803–2815.

612 [31] D. Kartsaklis, M. T. Pilehvar, and N. Collier, Mapping text to knowledge graph entities using multi-sense LSTMs, Proc.
613   2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, (2020), 1959–1970.

614 [32] J. Flisar and V. Podgorelec, Document enrichment using dbpedia ontology for short text classification, In Proc. of the 8th
615   International Conf. on Web Intelligence, Mining and Semantics, (2018), 1-9.

616 [33] H. Sun, T. Bedrax-Weiss, and W. W. Cohen, PullNet: Open domain question answering with iterative retrieval on
617   knowledge bases and text," Proc. of the 2019 Conf. on Empirical Methods in Natural Language Proc. and the 9th
618   International Joint Conf. on Natural Language Proc. (2019), 2380–2390.

619 [34] A. H. Miller, A. Fisch, J. Dodge, A. H. Karimi, A. Bordes, and J. Weston, Key-value memory networks for directly reading
620   documents, EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc., (2016), 1400–1409.

621 [35] M. Mountantonakis and Y. Tzitzikas, Large-scale semantic integration of linked data: A survey, ACM Computing Surveys
622   (CSUR), 52 (5) (2019), 1-40.

623 [36] C. G. Schuetz, S. Schausberger, and M. Schrefl, Building an active semantic data warehouse for precision dairy farming,
624   J. Organ. Comput. Electron. Commer., 28 (2) (2018), 122–141.

625 [37] M. Mountantonakis and Y. Tzitzikas, Content-based Union and Complement Metrics for Dataset Search over RDF
626   Knowledge Graphs, J. Data Inf. Qual., 12 (2) (2020), 1-31.

627 [38] S. Gonz and C. G. Eduardo, "Comparing BERT against traditional machine learning text classification," arXiv preprint
628   arXiv:2005.13012 (2020).

629 [39] R. Das, S. Dhuliawala, M. Zaheer,L. Vilnis,I. Durugkar,A. Krishnamurthy, … A.McCallum. Go for a walk and arrive at
630   the answer: Reasoning over paths in knowledge bases using reinforcement learning, International Conf. on Learning
631   Representations. (2018).

632 [40] Y. Qiu, Y. Wang, X. Jin, and K. Zhang, Stepwise reasoning for multi-relation question answering over knowledge graph
633   with weak supervision, WSDM 2020 - Proc. 13th Int. Conf. Web Search Data Min., (2020), 474–482.

634 [41] Y. Zhang, H. Dai, Z. Kozareva, A. J. Smola, and L. Song, Variational Reasoning for Question Answering with Knowledge
635   GraphZhang, Thirty-Second AAAI Conf. Artif. Intell., (2018).

636 [42] M. Wasim, D. Waqar, and D. Usman, A Survey of Datasets for Biomedical Question Answering Systems, Int. J. Adv.
637   Comput. Sci. Appl., vol. 8 (7) (2017), 1–5.

638 [43] A. Bordes, S. Chopra, and J. Weston, "Large-scale Simple Question Answering with Memory Networks, arXiv preprint
639   arXiv:1506.02075 (2015).

640 [44] K. Xu, S. Reddy, Y. Feng, S. Huang, and D. Zhao, Question answering on freebase via relation extraction and textual
641   evidence, 54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap., 4 (2016), 2326–2336.

642 [45] J. Bao, N. Duan, Z. Yan, M. Zhou, and T. Zhao, Constraint-based question answering with knowledge graph, COLING
643   2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap., (2016), 2503–2514.

644 [46] A. Talmor and J. Berant, The web as a knowledge-base for answering complex questions, NAACL HLT 2018 - 2018
645   Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., 1 (2018), 641–651.

646 [47] B. Fu, Y. Qiu, C. Tang, Y. Li, H. Yu, and J. Sun, A survey on complex question answering over knowledge base: Recent
647   advances and challenges, arXiv, (1) (2020), 1–19.

648 [48] R. Das, M. Zaheer, S. Reddy, and A. McCallum, Question answering on knowledge bases and text using universal schema
649   and memory networks, ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., 2 (2017), 358–
650   365.

651 [49] R. Usbeck, R. H. Gusmita, M. Saleem, and A. C. N. Ngomo, 9th challenge on question answering over linked data (QALD-
652   9)," CEUR Workshop Proc., 2241 (2018), 58–64.

653 [50] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, A sufficient condition for convergences of adam and rmsprop, Proc. IEEE

654   Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June (1) (2019), 11119–11127.

655   [51]   B. J. Erickson, P. Korfiatis, Z. Akkus, T. Kline, and K. Philbrick, Toolkits and Libraries for Deep Learning, J. Digit.
656   Imaging, 30 (4) (2017), 400–405.

657   [52]   F. T. Al-Khawaldeh, Hierarchical Attention Generative Adversarial Networks for Biomedical Texts Uncertainty
658   Detection, Int. J. Adv. Stud. Comput. Sci. Eng. Ijascse, 8 (6) (2019), 1–12.

659   [53]   M. Ulinski, S. Benjamin, and J. Hirschberg, Using Hedge Detection to Improve Committed Belief Tagging, (2018), 1–5.

660

Figure 1

Top: property-based relations and bottom: standard connections such as subsumption (e.g. subclass-of) and mereology (e.g., part-of).

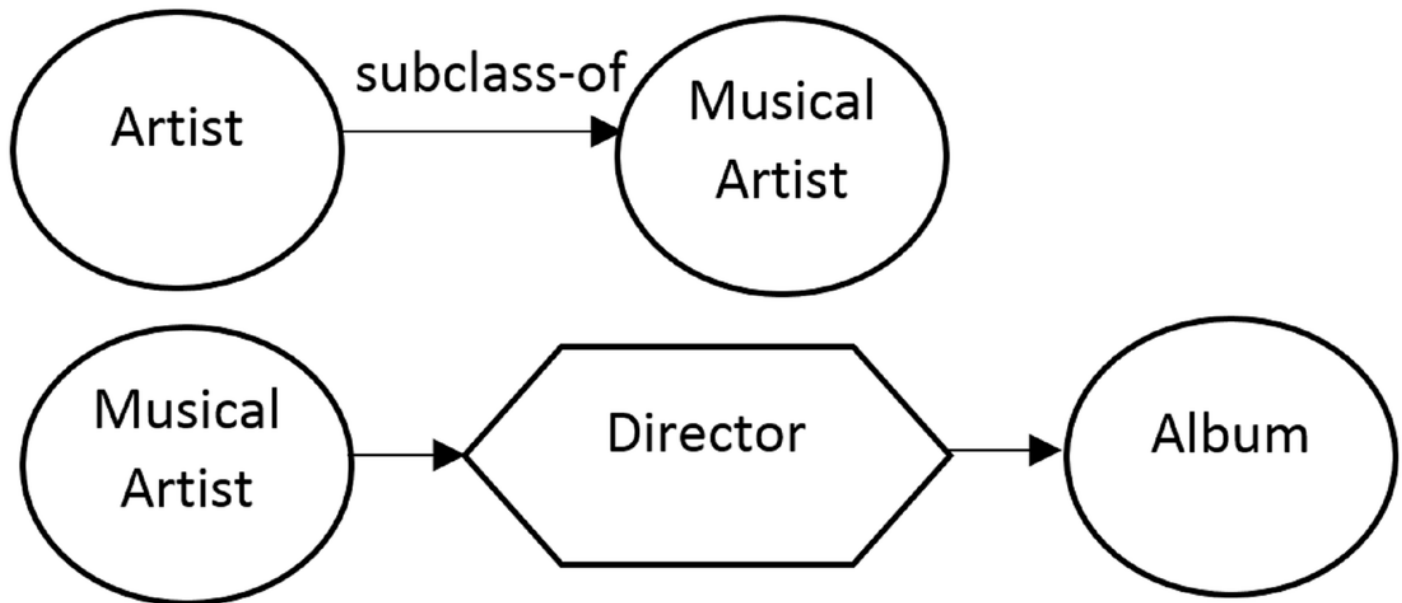



Figure 2

G) a piece of RDF graph and S) related schema for the triple in the dashed line
rectangle.

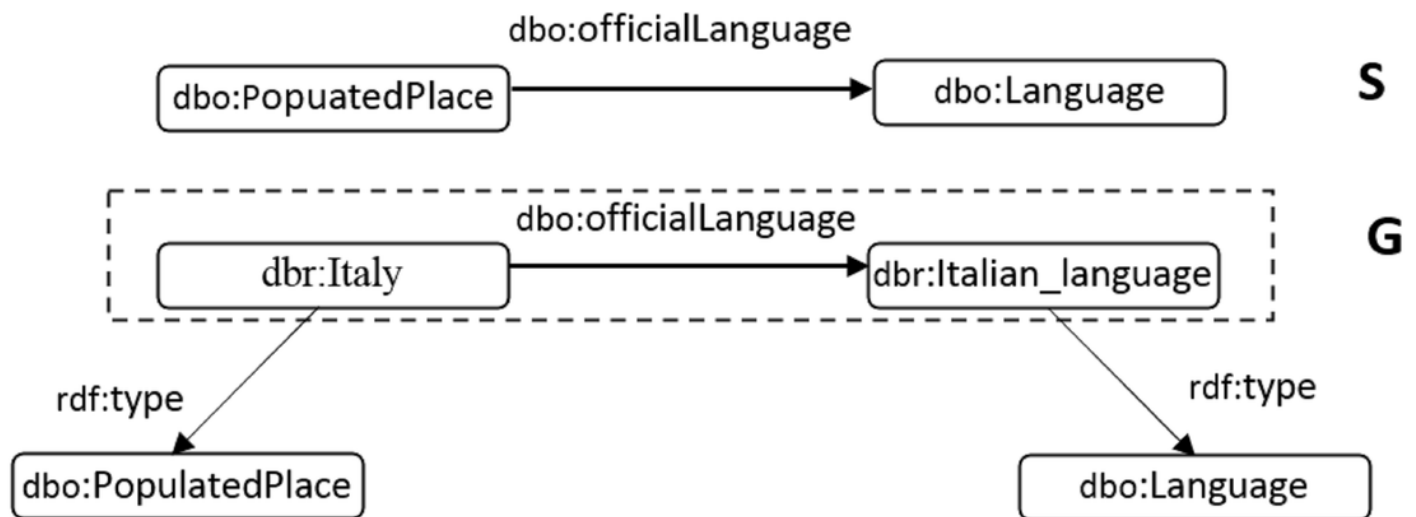



Figure 3

LODsyndesis indices [14]

Entity-based Triple Index

Entity	Property	Value	Dataset IDs
E1_Greco	LivedAt	Spain	4
	Carried out by	View Of Toledo	1,2
Heraklion	took place at	El-Greco-Birth	1
	population	15,1324	1,4

Entity Index

Entity	Dataset IDs
El_Greco	1,2,4
Heraklion	1

Property Index

Property	Dataset IDs
took place at	1,2
LivedAt	3,4

Literal Index

Literal	Dataset IDs
151,324	1,4
1542	2

Class Index

Class	Dataset IDs
Person	1,2
production	3

Figure 4

The architecture of LODQA [21]

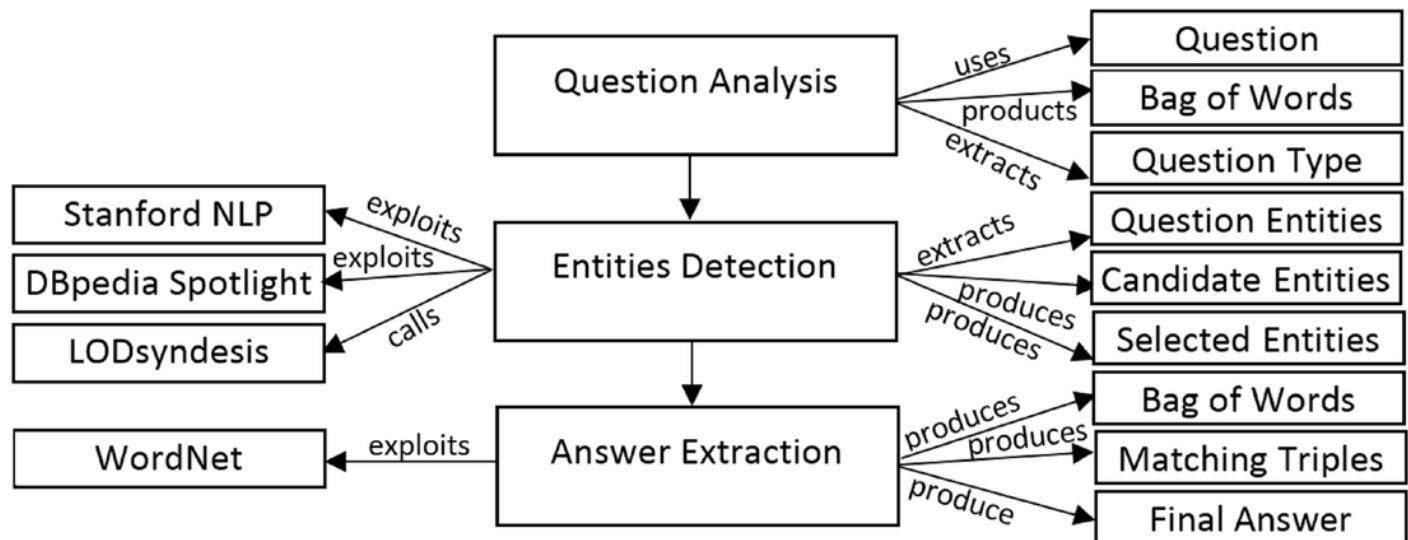



Figure 5

BERT-based KB Question Answering system (bertQA) [23]

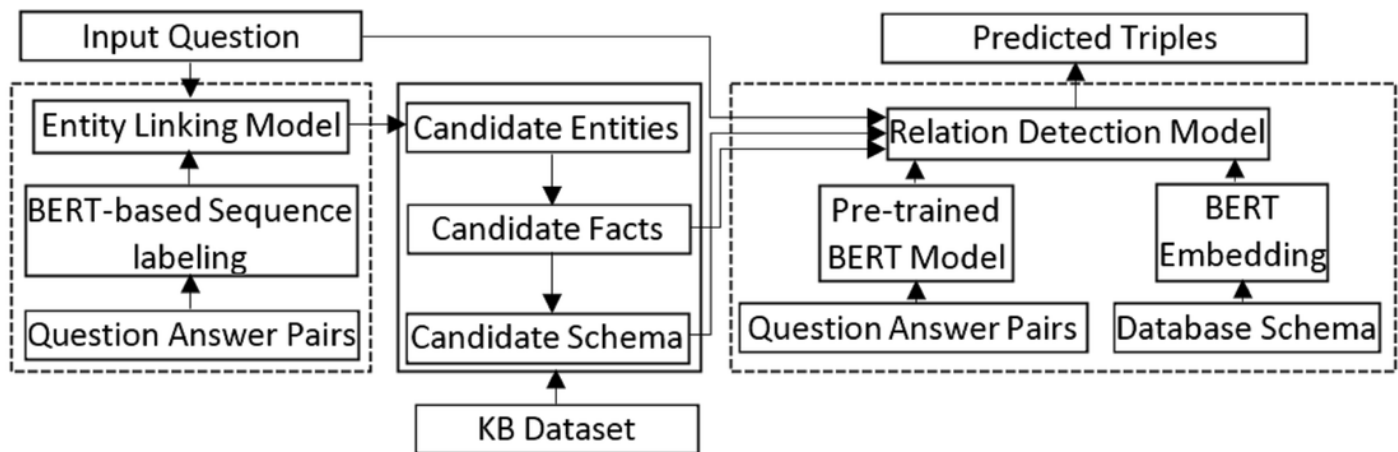



Figure 6

The proposed system, hybridSP, architecture for source prediction (left blue box) in the response extraction process.

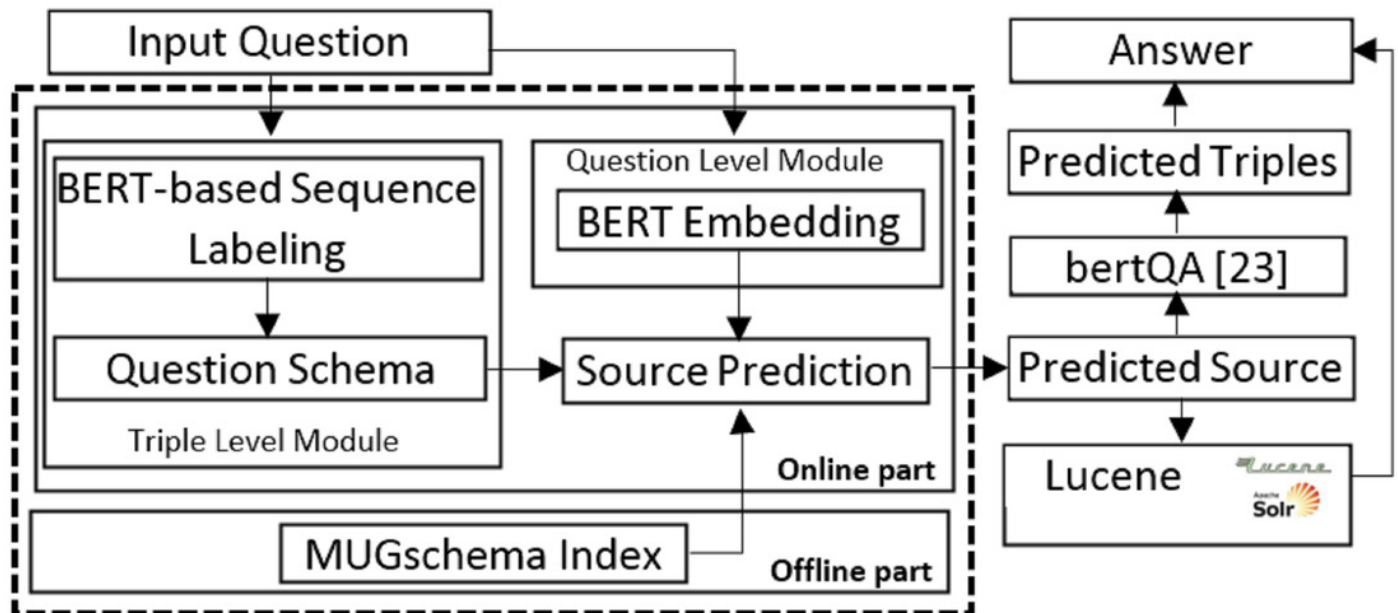



Figure 7

Offline part steps

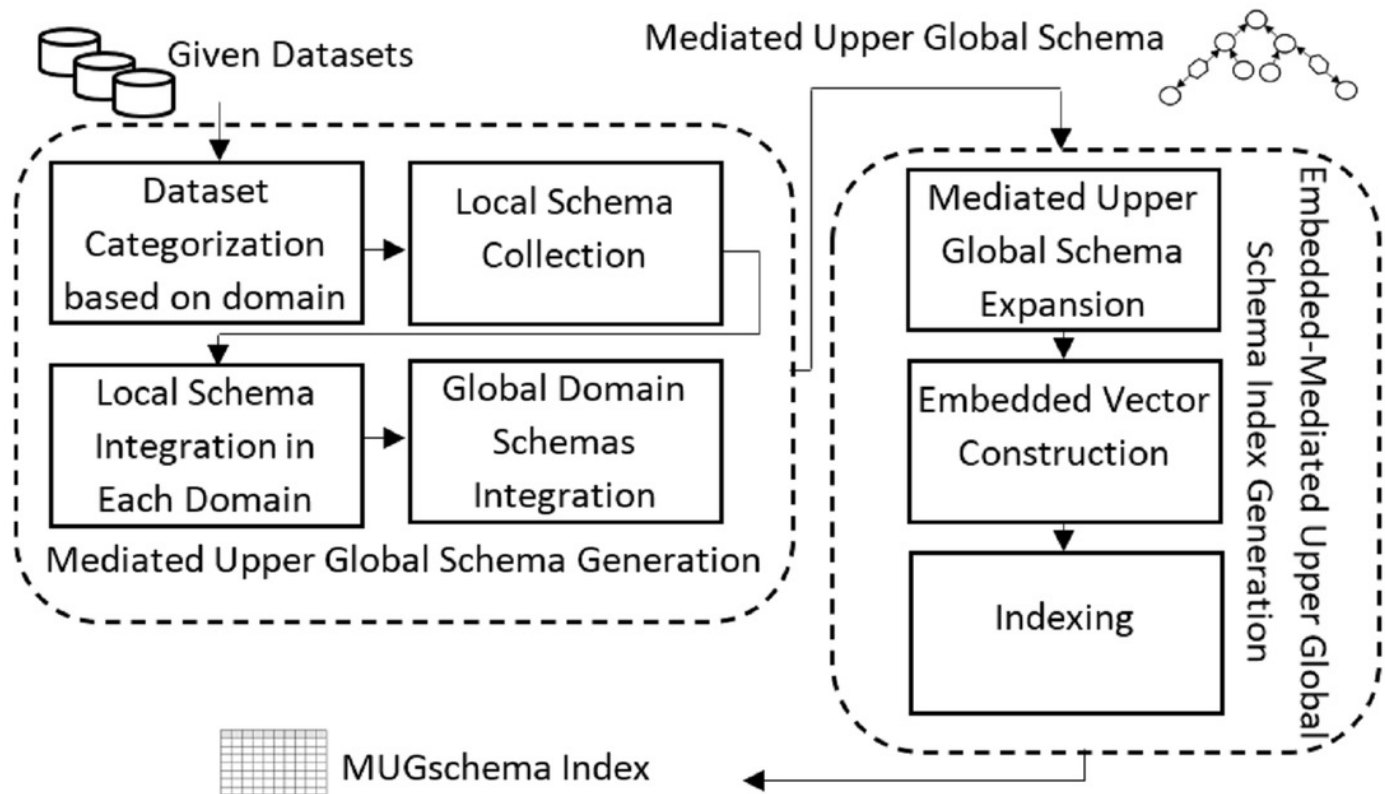



Figure 8

The mediated upper global schema, MUGschema, generation process.

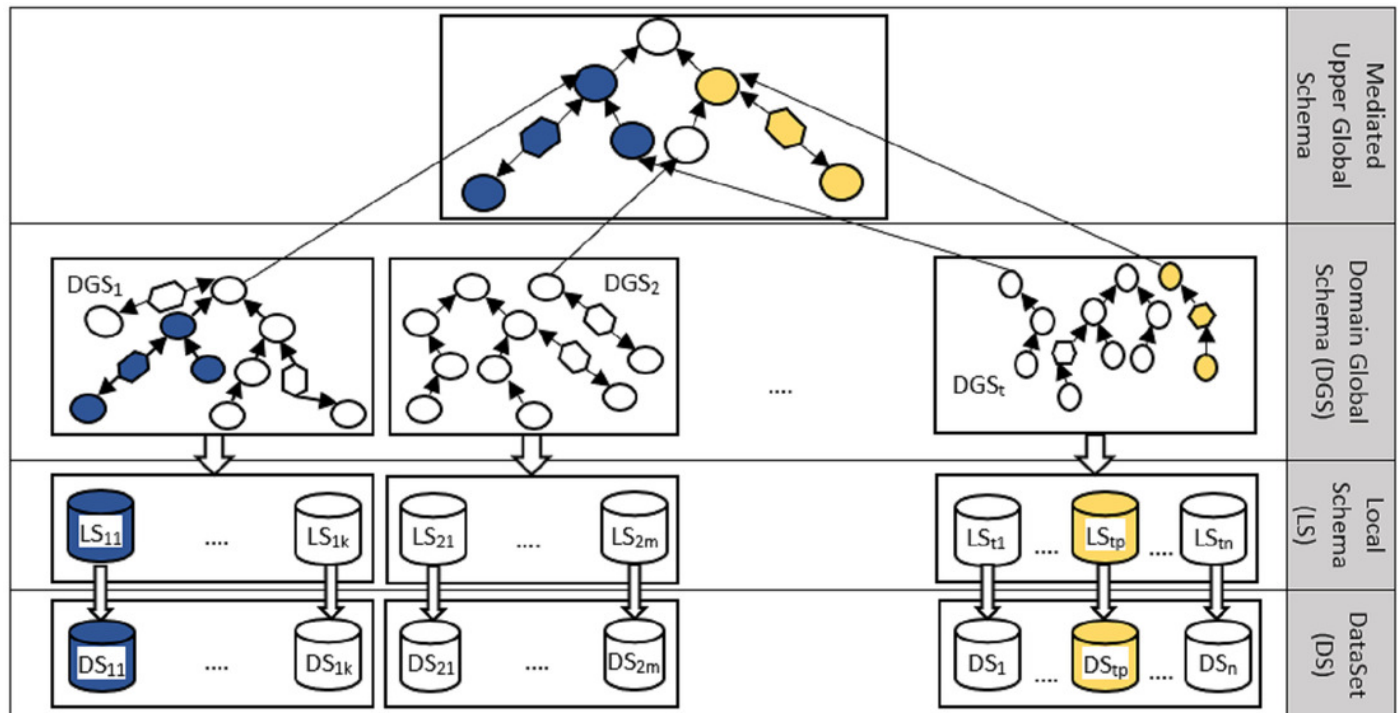



Figure 9

An Example to illustrate the mediated upper global level schema considering three
biomedical datasets Drugbank, Diseasome and Sider.

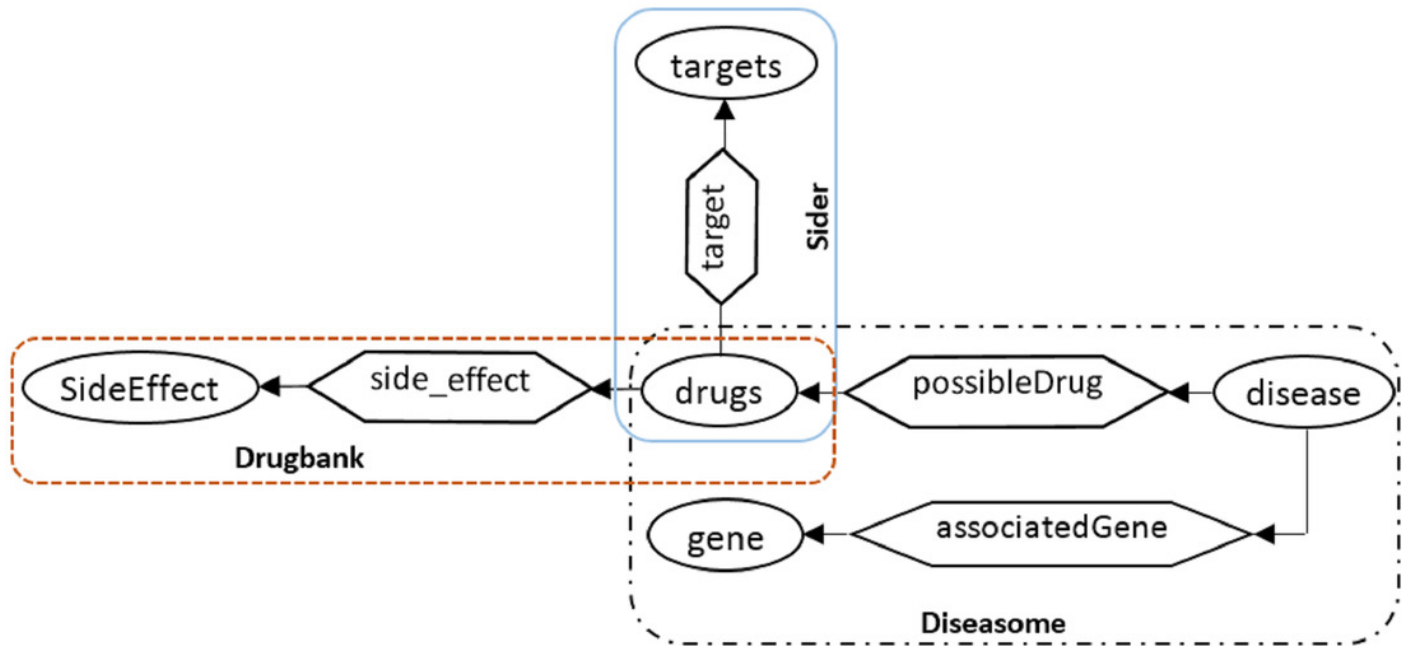



Figure 10

An Example to illustrate the triple expansion using neighbors.

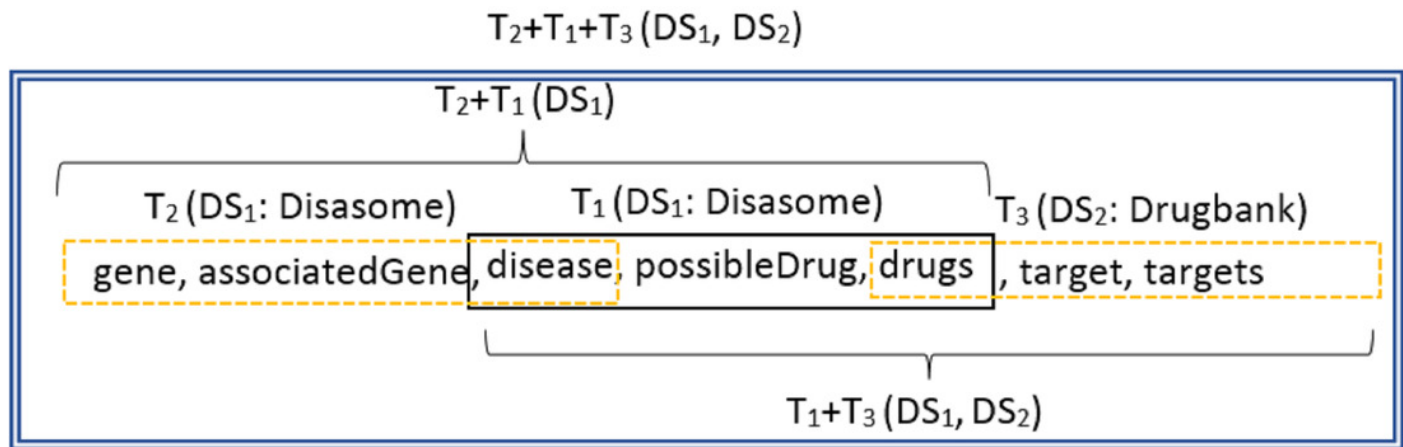



Figure 11

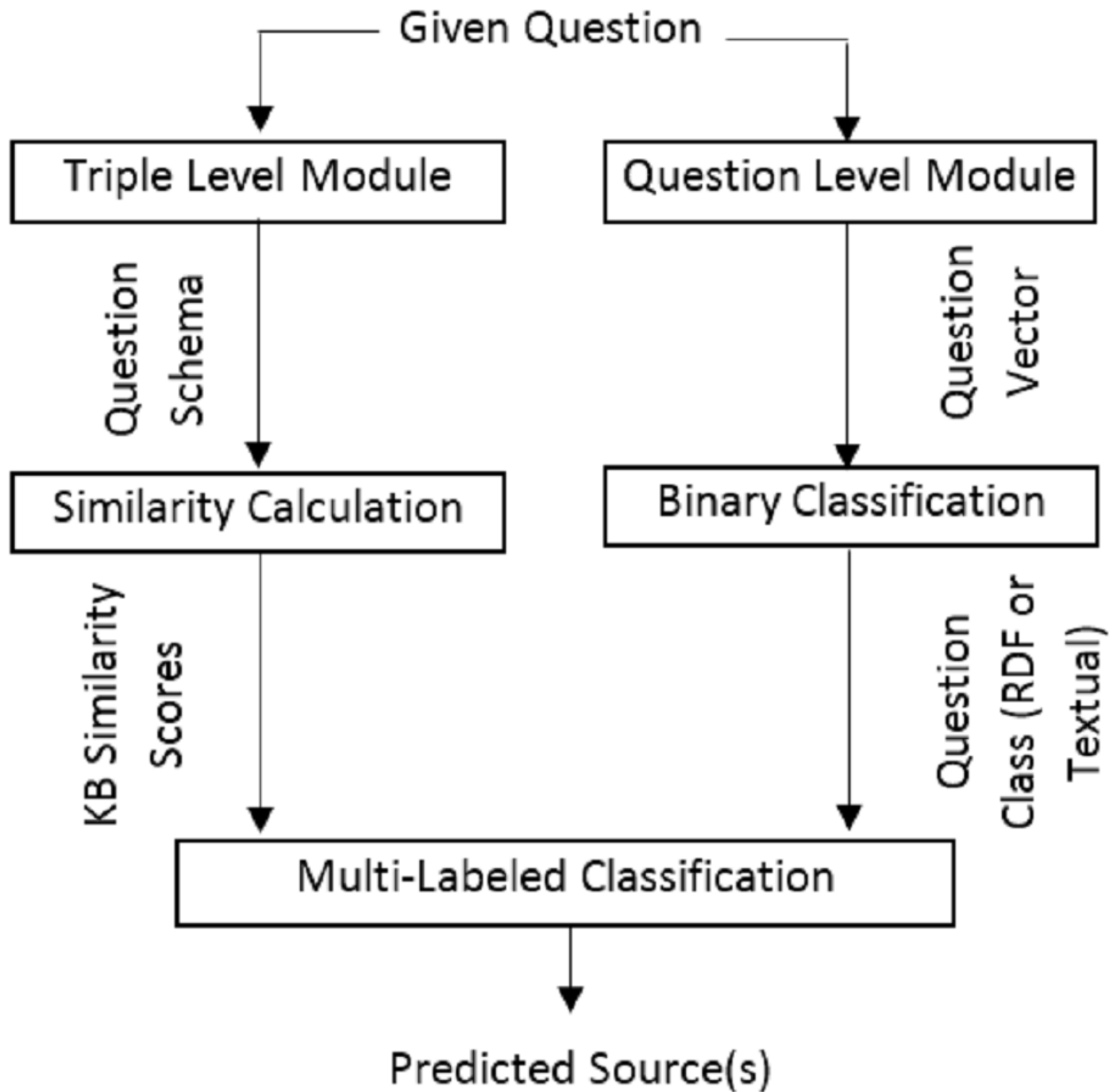
The first variation of hybridSP in online section to predict the answer source (s).

Predicted Source(s)

Figure 12

The second variation of hybridSP in online section to predict the answer source (s).

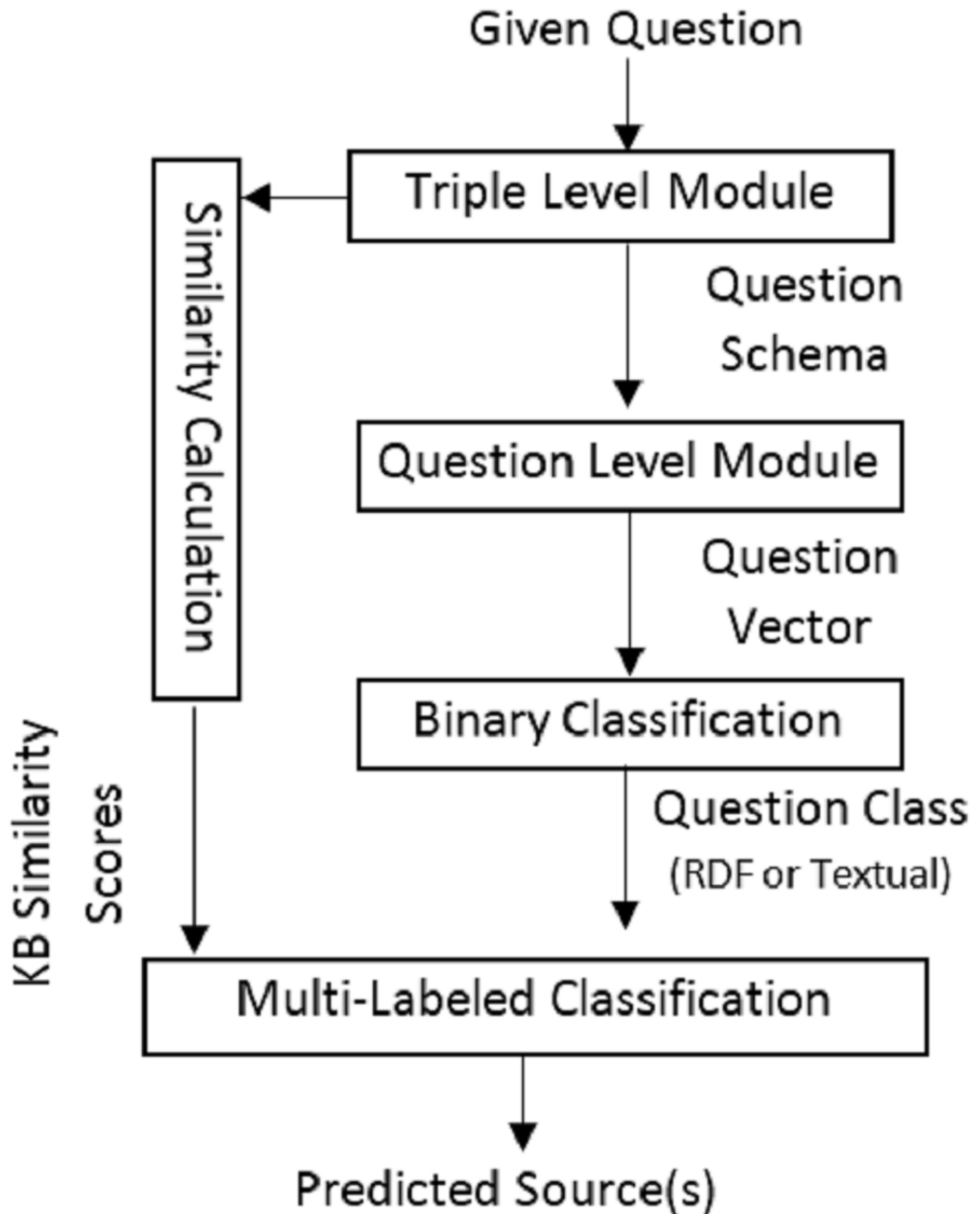



Figure 13

The proposed question schema generation model in online section.

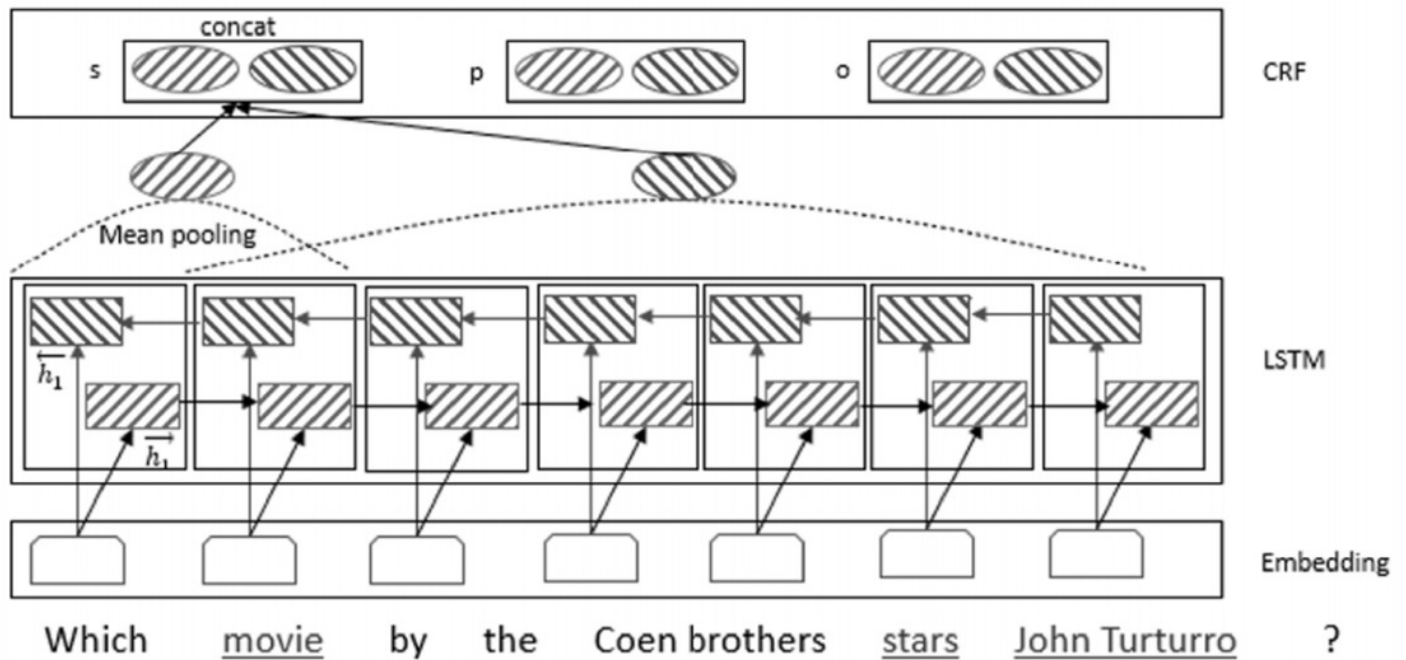



Figure 14

Comparison of accuracy metric among several benchmarks against different applied settings in question schema extraction. Denotations *lstm*, *crf*, *we*, and *pose* indicate use of the LTSM approach, CRF approach, word embeddings and P

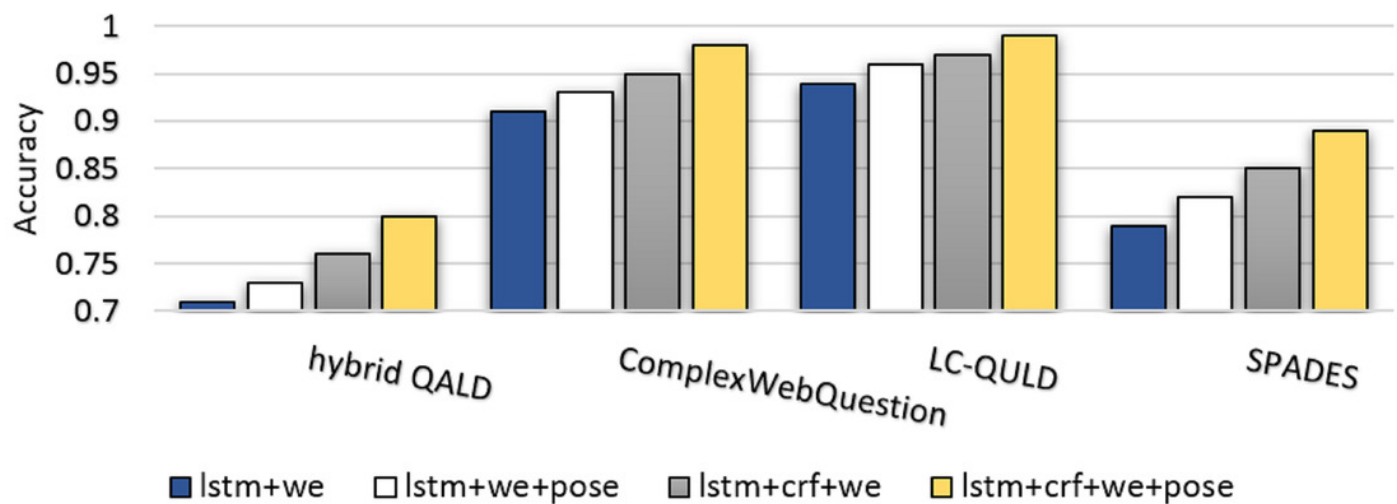



Figure 15

The accuracy for classifying questions in the KG or text classes. Results are calculated by applying the first variation (*FV*), second variation (*SV*), embedding feature vectors (*EF*), language feature vectors (*LF*), BERT and SVM class

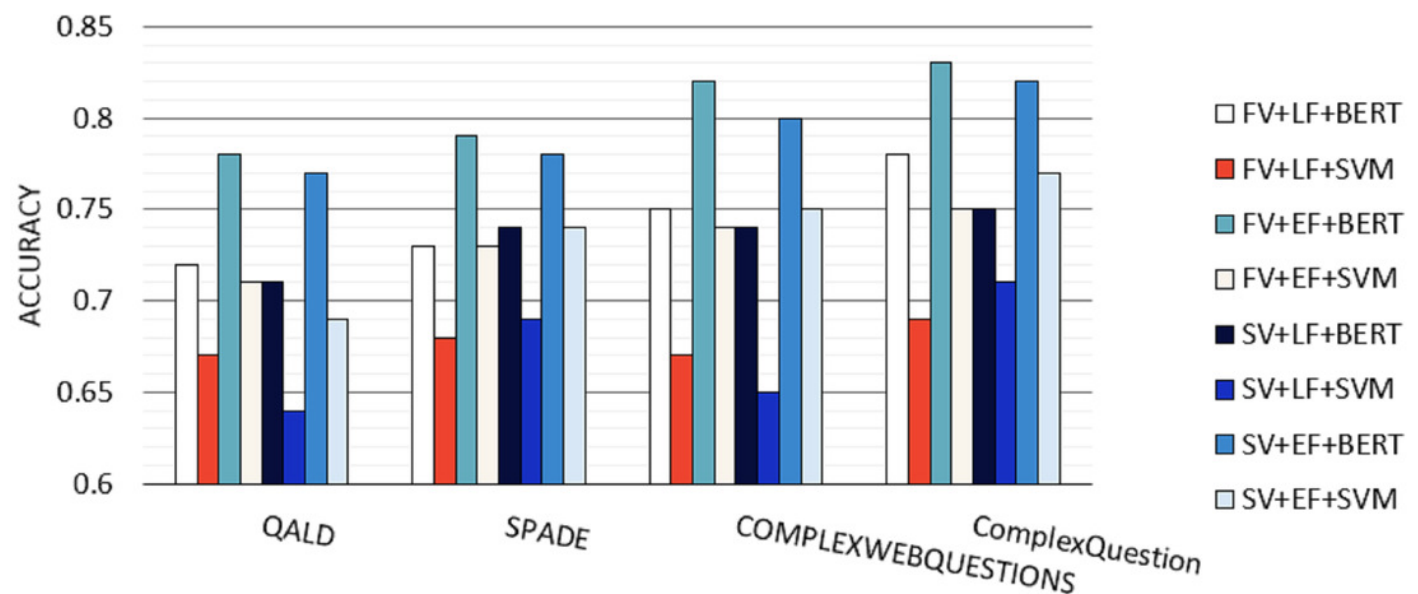



Figure 16

Comparison of hybridSP and LODsyndesis based on the accuracy-prediction time trade-off

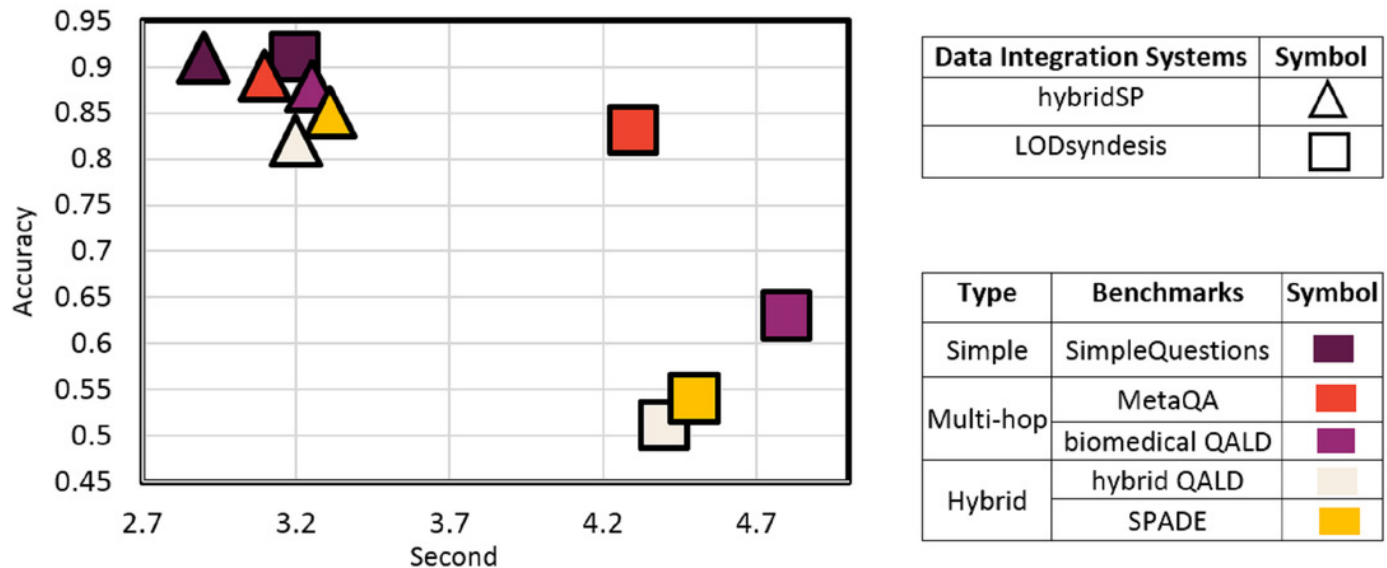
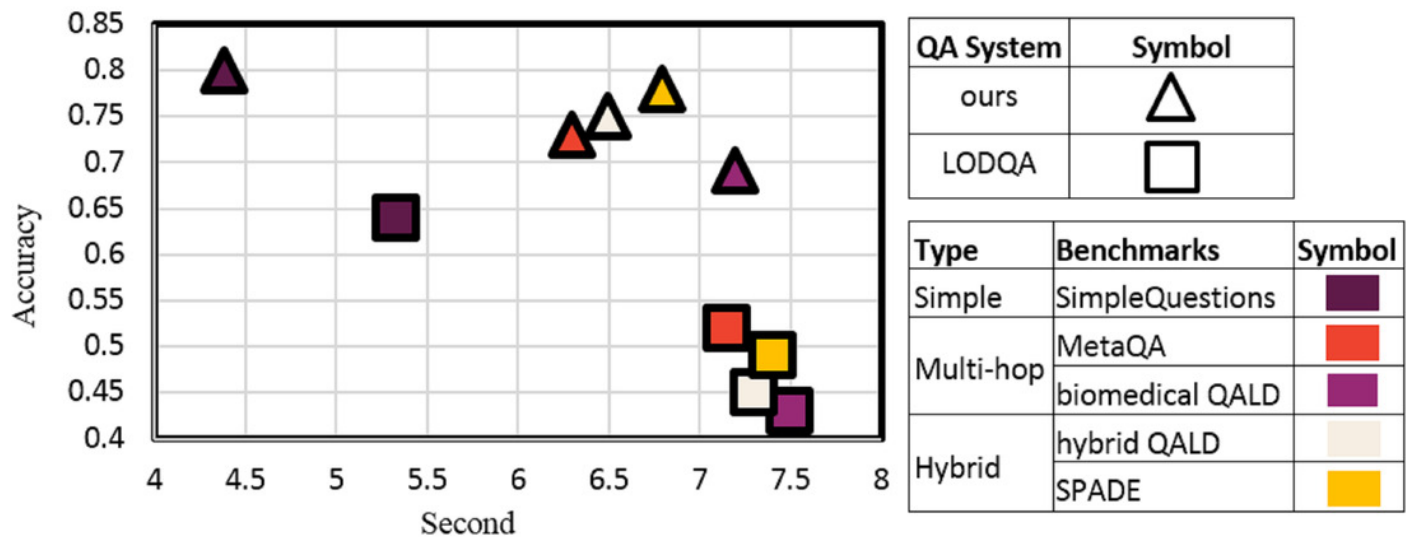



Figure 17

Comparison of BERT-based QA [23] and hybridSP method integration with LODQA [21] using LODsyndesis [14] based on the accuracy-response time trade-off

Examples of questions. sq_i and sq_i-triple stand for i-th sub-question and related its triple pattern, respectively.

Q_1= "Who composed the music for the film that depicts the early life of Jane Austen?"

sq_1=" What film depicts the early life of Jane Austen?

sq_1- triple= (? film, depicts, the early life of Jane Austen)

film= "Becoming Jane"

sq_2=" Who composed the music for the film?"

sq_2-triple= (Who, composed the music, film)

answer=" Adrian_Johnston".

Q_2= "Which writers had influenced the philosopher that refused a Nobel Prize?"

sq_1=" Which philosopher refused a Nobel Prize?"

sq_1- triple= (? philosopher, refused, Nobel Prize)

philosopher= "Jean-Paul Sartre"

sq_2=" Which writers had influenced the philosopher?"

sq_2-triple= (? writers, had influenced, philosopher)

answer=" Gustave_Flaubert".

Examples of simple question (1-hop) from SimpleQuestions [43] and 3-hop questions from the MetaQA [41] and biomedical QALD [42] datasest along with the triples needed to inference the relevant sources containing the answer.

Q₁: Which forest is Fires Creek in?   **T1**: (fires creek, containedby, nantahala national forest) – Freebase KB
Q₂: what types are the films directed by the director of "For Love or Money"?   **T₁**: (genre, type, film) - Freebase KB.   **T₂**: (Film, directedBy, person) - Freebase KB.   **T₃**: (person, director, movie) - Freebase KB.   **T₁+ T₂+ T₃**: (genre, type, film, Film, directedBy, person, person, director, movie) - Freebase KB.
Q₃: which genes are associated with diseases whose possible drugs target Cubilin?   **T₁**: (gene, associatedGene, disease) - Diseasome KB.   **T₂**: (disease, possible drugs, drugs) - Diseasome KB.   **T₃**: (drugs, target, targets) - Sider KB.   **T₁+ T₂+ T₃**: (gene, associatedGene, disease, possible drugs, drugs, target, targets) - Diseasome and Sider KBs.

Table 3_(on next page)

A piece of generated index for MUGschema index based on extended triples in Figure

Dataset ID	Embedded triple schema
1	$\vec{T}_1$
1	$\vec{T}_2$
2	$\vec{T}_3$
1	$\vec{T}_2 + \vec{T}_1$
2	$\vec{T}_1 + \vec{T}_3$
1,2	$\vec{T}_2 + \vec{T}_1 + \vec{T}_3$

General characteristics of benchmarks used in the evaluations.

Benchmark	Size	Type of Background	Source Background
SimpleQuestions [43]	1000	KB	FreeBase
LC-QULD [47]	5000	KB	DBpedia
hybrid QALD [49]	550	Hybrid	DBpedia, Wikipedia
biomedical QALD [42]	50	KB	Sider, Diseasome, Drugbank
ComplexWebQuestions [46]	34689	KB	FreeBase
MetaQA [41]	39282	KB	WikiMovies [34]
SPADES [48]	43000	Hybrid	FreeBase-CluWeb

Categories found by clustering question embedding vectors from the text class.

Cluster	Examples Containing Words
Sequence Adverbs	first, second, after, former
Adverbials of Location	around, front, among
Event Indicators of Time	during, when, the time of
Quality Adjectives	most, largest, deeper
Adverbs of Time	early, still, yet, already
Quantifiers	only, some, all
Adverbs of Frequency	often, sometimes, usually
Hedges	probably, definitely, certainly