

Multilingual Question Answering over Linked Data (QALD-3): lab overview

Philipp Cimiano¹, Vanessa Lopez², Christina Unger¹, Elena Cabrio³,
Axel-Cyrille Ngonga Ngomo⁴, and Sebastian Walter¹

¹ CITEC, Universität Bielefeld, Germany
{cimiano,cunger}@cit-ec.uni-bielefeld.de;

swalter@techfak.uni-bielefeld.de

² IBM Research, Dublin, Ireland
vanlopez@ie.ibm.com

³ INRIA Sophia-Antipolis, France
elena.cabrio@inria.fr

⁴ Universität Leipzig, Germany
ngonga@informatik.uni-leipzig.de

Abstract. The third instalment of the open challenge on Question Answering over Linked Data (QALD-3) has been conducted as a half-day lab at CLEF 2013. Differently from previous editions of the challenge, QALD-3 put a strong emphasis on multilinguality, offering two tasks: one on multilingual question answering and one on ontology lexicalization. While no submissions were received for the latter, the former attracted six teams who submitted their systems' results on the provided datasets. This paper provides an overview of QALD-3, discussing the approaches experimented by the participating systems as well as the obtained results.

1 Introduction

While more and more semantic data is published on the web, the question of how typical web users can access this body of knowledge becomes of crucial importance. Over the past years, there is a growing amount of research on interaction paradigms that allow end users to profit from the expressive power of Semantic Web standards while at the same time hiding their complexity behind an intuitive and easy-to-use interface; for an overview see [11]. Especially natural language interfaces have received wide attention, as they allow users to express arbitrarily complex information needs in an intuitive fashion and, at least in principle, in their own language. The key challenge lies in translating the users' information needs into a form such that they can be evaluated using standard Semantic Web query processing and inferencing techniques. To this end, systems have to deal with a heterogeneous, distributed and very large set of highly interconnected data. The availability of such an amount of open and structured data has no precedents in computer science and approaches that can deal with the specific character of linked data are urgently needed. In addition, multilinguality has become an issue of major interest for the Semantic Web community,

as both the number of actors creating and publishing data in languages other than English, as well as the amount of users that access this data and speak native languages other than English is growing substantially. In order to achieve the goal that users from all countries have access to the same information, there is an impending need for systems that can help in overcoming language barriers by facilitating multilingual access to semantic data originally produced for a different culture and language.

The main objective of the open challenges on *question answering over linked data*¹ (QALD) is to provide an up-to-date, demanding benchmark that establishes a standard against which question answering systems over structured data can be evaluated and compared. QALD-3 is the third instalment of the QALD open challenge, organized as a half-day lab at CLEF 2013.

The rest of the paper describes the previous editions of the challenge (Section 2), details the main novelties and the experimental setting of QALD-3 (Section 3) and the results obtained by the participating systems (Section 4). Section 5) then draws some conclusions about the current edition and proposes ideas for next editions of the challenge.

2 Previous QALD challenges

The QALD challenges aim to bring together researchers and developers from different communities, including NLP, Semantic Web, human-computer interaction, and databases. The first edition, QALD-1, was organised in the context of the workshop *Question Answering Over Linked Data* at ESWC 2011. The second edition, QALD-2, was run in the context of the workshop *Interacting With Linked Data* at ESWC 2012 and broadened the scope to also include other paradigms for interacting with linked data as well as encourage communication across interaction paradigms.

In the context of QALD-1, two datasets were made available—DBpedia and an RDF export of the MusicBrainz database—together with a set of 50 training and 50 test questions each. These questions were created by a student assistant with no background in question answering in order to avoid a bias towards a particular approach. The questions were designed to present potential user questions and to include a wide range of challenges such as lexical ambiguities and complex syntactical structures. All training questions were annotated with corresponding SPARQL queries. For QALD-2, both question sets were combined to build a new training set, and a newly created test set was provided, leading to 100 training and 100 test questions for DBpedia, and 100 training and 50 test questions for MusicBrainz. In addition, a few out-of-scope questions were added to each question set, i.e., questions to which the datasets do not contain the answer, in order to test the ability of participating systems to judge whether a failure to provide an answer lies in the dataset or in the system itself. Further, we provided a small set of questions that could only be answered by combining

¹ <http://www.sc.cit-ec.uni-bielefeld.de/qald>

information from both datasets, DBpedia and MusicBrainz, thus testing a system’s ability to combine several linked information sources when searching for an answer. All QALD-2 questions were additionally annotated with keywords in order to encourage keyword-based approaches to take part in the challenge.

For a detailed description of the challenge as well as the participating systems and their results, see [10].

3 QALD-3

Capitalizing on the positive feedback which QALD has received from the Semantic Web and NLP communities, the third challenge was decided to make a step forward by introducing new elements. To this end, QALD-3 proposed two separate tasks: *multilingual question answering*, that keeps the basic structure of the previous challenges unchanged but introduces multilingualism as the major innovation, and *ontology lexicalization*, aimed at all methods that (semi-)automatically create lexicalizations of ontology concepts. In the following, we present more details about the proposed tasks and the resources we made available to the participants.

3.1 Task 1: Multilingual question answering

Task 1 aims at all question answering systems that mediate between a user, expressing his or her information need in natural language, and semantic data. Given a RDF dataset and a natural language question or set of keywords in one of six languages (English, Spanish, German, Italian, French, Dutch), the participating systems had to return either the correct answers, or a SPARQL query that retrieves these answers. In order to evaluate and compare participating systems, three RDF datasets were provided:

- English DBpedia 3.8² (including links, most importantly to YAGO categories³ and MusicBrainz⁴), a community effort to extract structured information from Wikipedia and to make this information available as RDF data
- Spanish DBpedia⁵, containing information from Wikipedia extracted in Spanish (containing almost 100 million RDF triples)
- MusicBrainz, a collaborative effort to create an open content music database. The dataset provided for the challenge is an RDF export containing all classes (artists, albums and tracks) and the most important properties of the MusicBrainz database

These datasets could either be downloaded or accessed through a provided SPARQL endpoint.

² <http://dbpedia.org>

³ <http://www.mpi-inf.mpg.de/yago-naga/yago/>

⁴ musicbrainz.org

⁵ <http://es.dbpedia.org>

To get acquainted with the datasets and possible questions, a set of 100 training questions for each dataset (i.e. English DBpedia, Spanish DBpedia and MusicBrainz) was provided. Later, systems were evaluated on 100 different test questions. Both training and test questions were mainly adopted from the QALD-2 challenge, slightly modified in order to account for changes in the DBpedia dataset and in order to include feedback obtained from participants of the first two challenges. As major innovation, all questions and keywords were translated into six different languages: English, Spanish, German, Italian, French, and Dutch. Here are some English example questions from the training sets:

- *DBpedia*:
 - 5 How many monarchical countries are there in Europe?
 - 58 Who produced the most films?
 - 74 Which capitals in Europe were host cities of the summer Olympic games?
 - 85 In which films did Julia Roberts as well as Richard Gere play?
- *Spanish DBpedia*:
 - 2 Who was the son of Alfonso López Pumarejo married to?
 - 4 In which city did Eva Perón die?
 - 20 What is the area code of Barcelona?
 - 40 How many films did Pedro Almodóvar produce?
- *MusicBrainz*:
 - 2 Which groups was David Bowie a member of?
 - 44 How many versions of the song Smells Like Teen Spirit are there?
 - 79 Who did the vocals on the album Sabotage?
 - 89 When were The Vertigos founded?

All training and test questions were manually annotated with keywords, corresponding SPARQL queries and with answers retrieved from the provided SPARQL endpoint. Annotations were provided in an XML format. Each of the questions specifies an ID for the question together with a range of other attributes explained below, the natural language string of the question in the six languages, keywords in the same languages, a corresponding SPARQL query, as well as the answers this query returns. Along with a unique ID, the following attributes were specified for each question:

- **answertype** gives the answer type, which can be one of the following: **resource** (one or many resources, for which the URI is provided), **string** (a string value), **number** (a numerical value such as 47 or 1.8), **date** (a date provided in the format YYYY-MM-DD, e.g. 1983-11-02), **boolean** (either **true** or **false**)
- **aggregation** indicates whether any operations beyond triple pattern matching are required to answer the question (e.g., counting, filters, ordering)
- **onlydbo** is given only for DBpedia questions and reports whether the query relies solely on concepts from the DBpedia ontology

Here is an example from the DBpedia training set:

```
1 <question id="36" answertype="resource"
2           aggregation="false" onlydbo="false">
3
4 <string lang="en">
5 Through which countries does the Yenisei river flow?
6 </string>
7 <string lang="de">
8 Durch welche Länder fließt der Yenisei?
9 </string>
10 <string lang="es">
11 ¿Por qué países fluye el río Yenisei?
12 </string>
13 <string lang="it">
14 Attraverso quali stati scorre il fiume Yenisei?
15 </string>
16 <string lang="fr">
17 Quels sont les pays traversés par l'Iénisseï?
18 </string>
19 <string lang="nl">
20 Door welke landen stroomt de Jenisej?
21 </string>
22
23 <keywords lang=en>
24 Yenisei river, flow through, country
25 </keywords>
26 ...
27
28 <query>
29 PREFIX res: <http://dbpedia.org/resource/>
30 PREFIX dbp: <http://dbpedia.org/property/>
31 SELECT DISTINCT ?uri WHERE {
32   res:Yenisei River dbp:country ?uri .
33 }
34 </query>
35
36 <answers>
37 <answer>
38 <uri>http://dbpedia.org/resource/Mongolia</uri>
39 </answer>
40 <answer>
41 <uri>http://dbpedia.org/resource/Russia</uri>
42 </answer>
43 </answers>
44 </question>
```

As an additional challenge, some of the training and test questions are out of scope, i.e. they cannot be answered with respect to the dataset.

3.2 Task 2: ontology lexicalization

Multilingual information access can be facilitated by the availability of lexica in different languages, for example allowing for an easy mapping of Spanish, German, and French natural language expressions to English ontology labels. The task consisted in finding English lexicalizations of a set of classes and properties from the DBpedia ontology, for example in a Wikipedia corpus. The training data provided to the participating systems consisted of a set of 10 classes and 30 properties from the DBpedia ontology, as well as a lexicon containing lexicalizations of those classes and properties in *lemon*⁶ format. Classes and properties were randomly chosen from the DBpedia ontology (properties with less than 20 entity pairs to properties with over 100,000 entity pairs). Here is an example of expected lexicalizations for the DBpedia class `TradeUnion`:

```
1 :TradeUnion a lemon:LexicalEntry ;
2   lemon:canonicalForm [ lemon:writtenRep "trade union"@en ] ;
3   lemon:sense [ lemon:reference
4                 <http://dbpedia.org/ontology/TradeUnion> ] ;
5   lexinfo:partOfSpeech lexinfo:noun .
6
7 :LaborUnion a lemon:LexicalEntry ;
8   lemon:canonicalForm [ lemon:writtenRep "labor union"@en ] ;
9   lemon:sense [ lemon:reference
10                 <http://dbpedia.org/ontology/TradeUnion> ] ;
11  lexinfo:partOfSpeech lexinfo:noun .
```

3.3 Evaluation measures

The results submitted by participating systems were automatically compared to the gold standard results.

Task 1 For each question q , precision, recall and F-measure were computed as follows:

$$\text{Recall}(q) = \frac{\text{number of correct system answers for } q}{\text{number of gold standard answers for } q}$$
$$\text{Precision}(q) = \frac{\text{number of correct system answers for } q}{\text{number of system answers for } q}$$

$$\text{F-Measure}(q) = \frac{2 * \text{Precision}(q) \times \text{Recall}(q)}{\text{Precision}(q) + \text{Recall}(q)}$$

On the basis of these measures, overall precision and recall values as well as an overall F-measure value were computed as the average mean of the precision, recall and F-measure values for all questions. In the results reported in Section 4 below, precision, recall and F-measure values refer to the averaged values.

⁶ <http://lemon-model.net>

Task 2 For each property, the uploaded lexical entries were evaluated automatically by comparing them to the manually created lexical entries along two dimensions: i) lexical precision, lexical recall and lexical F-measure, and ii) lexical accuracy. The first dimension evaluates how many of the gold standard entries for a property were submitted by the participants, and how many of the automatically generated entries are among the gold standard entries (precision), where two entries count as the same lexicalization if their lemma, part of speech and sense coincide. Thus lexical precision P_{lex} and recall R_{lex} for a property p are defined as follows:

$$P_{lex}(p) = \frac{|entries_{auto}(p) \cap entries_{gold}(p)|}{|entries_{auto}(p)|}$$

$$R_{lex}(p) = \frac{|entries_{auto}(p) \cap entries_{gold}(p)|}{|entries_{gold}(p)|}$$

where $entries_{auto}(p)$ is the set of entries for the property p in the automatically constructed lexicon, while $entries_{gold}(p)$ is the set of entries for the property p in the manually constructed gold lexicon. The F-measure $F_{lex}(p)$ is then defined as the harmonic mean of $P_{lex}(p)$ and $R_{lex}(p)$, as usual.

The second dimension, lexical accuracy, is necessary in order to evaluate whether the specified subcategorization frame and its arguments are correct, and whether these syntactic arguments have been mapped correctly to the semantic arguments (domain and range) of the property in question. The accuracy of an automatically generated lexical entry l_{auto} for a property p w.r.t. the corresponding gold standard entry l_{gold} is therefore defined as:

$$A_p(l_{auto}) = (frameEq(l_{auto}, l_{gold}) + \frac{|args(l_{auto}) \cap args(l_{gold})|}{|args(l_{gold})|} + \frac{\sum_{a \in args(l_{auto})} map(a)}{|args(l_{auto})|})/3$$

Where $frameEq(l_1, l_2)$ is 1 if the subcategorization frame of l_1 is the same as the subcategorization frame of l_2 , and 0 otherwise, where $args(l)$ returns the syntactic arguments of l 's frame, and where

$$map(a) = \begin{cases} 1, & \text{if } a \text{ in } l_{auto} \text{ has been mapped to the same semantic argument} \\ & \text{of } p \text{ as in } l_{gold} \\ 0, & \text{otherwise} \end{cases}$$

When comparing the argument mapping of the automatically generated entry with that of the gold standard entry, only the class of the argument is considered, i.e. *subject* or *object*. This abstracts from the specific type of subject (e.g. *copulative subject*) and object (e.g. *indirect object*, *prepositional object*, etc.) and therefore allows for an evaluation of the argument mappings independently of the correctness of the frame and frame arguments. The lexical accuracy $A_{lex}(p)$ for a property p is then computed as the average mean of the accuracy values of each generated lexicalization. All measures are computed for each property and then averaged for all properties.

4 Participating systems, results and discussion

Six teams participated in QALD-3, two groups more than in last year’s challenge: five teams from Europe (three from France, one from Germany and one from Italy), and one from Asia (China). Participants were allowed to submit runs to one or both of the tasks. Six participants took part in multilingual question answering task, five participants on the DBpedia track only (and all of them on English questions only), and one participant on both DBpedia and MusicBrainz. No runs were submitted for the ontology lexicalization task.

4.1 Participating systems

The participating systems follow different approaches to question answering over linked data. For question interpretation, some rely on linguistic strategies, e.g. the analysis of syntactic patterns, while others implement statistical approaches. In contrast to systems that take the provided natural language question as input, *squall2sparql* takes as input questions in SQUALL, a controlled natural language for English, and *Scalewelis* is based on faceted search instead of question interpretation. In the following, we give some details on the participating systems.

Intui2 [3] is a prototype system for question answering over linked data that can answer natural language questions with respect to a given RDF dataset by analyzing the questions in terms of the syntactic constituents (*synfragments*) they are composed of. Syntactically, a *synfragment* corresponds to a subtree of the syntactic parse tree of the question, and semantically, it is a minimal span of text that can be interpreted as a concept URI, an RDF triple or a complex RDF query. These *synfragments* are then compositionally combined to an interpretation of the whole input question.

SWIP [14] relies on the use of query patterns to address the task of interpreting natural language queries. The query interpretation process consists of two main steps. First, the natural language question is translated into a pivot query, capturing the query focus, a dependency analysis and the extracted relations between substrings of the natural language question. Second, predefined query patterns are mapped to the pivot query, obtaining a list of potential interpretations of the user question, which are then ranked according to their estimated relevance and proposed to the user in form of reformulated natural language questions.

CASIA [8] implements a pipeline consisting of question analysis, resource mapping and SPARQL generation. More specifically, the system first transforms and represents natural language questions as a set of *query triples* of the form $\langle \text{subject}, \text{predicate}, \text{object} \rangle$, based on a shallow and deep linguistic analysis. Second, it instantiates these query triples with corresponding resources from DBpedia, resulting in *ontology triples*. Third, based on the ontology triples and question type, SPARQL queries are constructed. Finally, the candidate queries are validated and ranked, and the best query is selected.

squall2sparql [5] is a translator from SQUALL, a controlled natural language for English, to SPARQL. Given a SQUALL sentence, the system first translates it into an intermediate logical representation based on Montague grammar. This intermediate representation is then translated into SPARQL by mapping logical constructs to combinations of SPARQL constructs.

*Scalewelis*⁷ [7] is a faceted search system that guides the user through the search for an answer. Starting from an initial SPARQL query, facets are created for the first 1,000 results retrieved by that query, consisting of the classes the results belong to as well as properties that relate the results to other entities in the dataset. The user's selection of a facet is then used to refine the query until the answer is found.

The *RTV* system [6] integrates lexical semantic modelling and statistical inferences within a complex architecture that decomposes the natural language interpretation task into a cascade of three different stages: i) the selection of salient information from the question (i.e. predicate, arguments and properties), ii) the location of the salient information in the ontology through joint disambiguation of all candidates, and iii) the compilation of the final query against RDF triples. This architecture exploits a Hidden Markov Model (HMM) to select the proper ontological triples according to the graph nature of RDF. In particular, for each query an HMM model is produced whose Viterbi solution is the comprehensive joint disambiguation across the sentence elements.

4.2 Used external resources and tools

Table 1 shows the external resources and tools exploited by participating systems. Among the resources, Wikipedia and WordNet are used for semantic knowledge extraction (e.g. for calculating similarity among words in Intui2). Concerning external tools, text processing tools are used for questions preprocessing (i.e. Stanford CoreNLP, MaltParser and Chaos), while information retrieval tools such as Lucene are used to index Wikipedia versions in the RTV system, or to obtain string similarity scores in SWIP.

Two of the participating systems do not rely on linguistic resources at all: *squall2sparql*, where the use of controlled English as input language bypasses most of the problems related to language variability, and *Scalewelis*, which relies on faceted search rather than question interpretation.

4.3 Results

Tables 2 and Table 3 report on the results obtained by the participating systems over DBpedia and MusicBrainz datasets, respectively. The column *processed* states for how many of the questions the system provided an answer, *right* specifies how many of these questions were answered with an F-measure of 1, and

⁷ <http://lisfs2008.irisa.fr/scalewelis/>

Table 1. External resources and tools used by the participating systems

<i>Resources</i>	CASIA	Intui2	SWIP	RTV
WordNet [4]	+	+	-	-
Wikipedia	-	+	-	+
PATTY ⁸ [12]	+	-	-	-
<i>Tools</i>	CASIA	Intui2	SWIP	RTV
WS4J (WordNet Similarity for Java) ⁹	-	+	-	-
Chaos parser [2]	-	-	-	+
MaltParser [13]	-	-	+	-
Stanford CoreNLP [9]	+	+	-	-
Jena ARQ query engine	-	+	-	-
Lucene ¹⁰	-	-	-	+
LARQ (Lucene + ARQ) ¹¹	-	-	+	-

partially specifies how many of the questions were answered with an F-measure strictly between 0 and 1. On the DBpedia dataset, the best F-measure was 0.9 and the lowest was 0.17, the average being 0.4. These results are comparable to the results achieved in earlier challenges, showing that the level of complexity of the questions is still very demanding.

Table 2. Results for DBpedia test set

<i>System</i>	<i>Total</i>	<i>Processed</i>	<i>Right</i>	<i>Partially</i>	<i>Recall</i>	<i>Precision</i>	<i>F-measure</i>
squall2sparql	99	99	80	13	0.88	0.93	0.90
CASIA	99	52	29	8	0.36	0.35	0.36
Scalewelis	99	70	32	1	0.33	0.33	0.33
RTV	99	55	30	4	0.34	0.32	0.33
Intui2	99	99	28	4	0.32	0.32	0.32
SWIP	99	21	15	2	0.16	0.17	0.17

Table 3. Results for MusicBrainz test set

<i>System</i>	<i>Total</i>	<i>Processed</i>	<i>Right</i>	<i>Partially</i>	<i>Recall</i>	<i>Precision</i>	<i>F-measure</i>
SWIP	50	33	24	2	0.51	0.51	0.51

The following questions on DBpedia were answered by all systems:

<i>ID</i>	<i>Question</i>
21	What is the capital of Canada?
22	Who is the governor of Wyoming?
30	What is the birth name of Angela Merkel?
68	How many employees does Google have?

And the following questions on DBpedia were answered by no systems:

ID Question

14 Give me all members of Prodigy.

16 Does the new Battlestar Galactica series have more episodes than the old one?

92 Show me all songs from Bruce Springsteen released between 1980 and 1990.

96 Give me all B-sides of the Ramones.

Of the questions in the test set, 45 queries require to search the answer using other namespaces than the DBpedia ontology (attribute `onlydbo=false`), such as YAGO or FOAF, and 19 queries require aggregation operations (attribute `aggregation=true`), such as comparisons, like in 16 above, superlatives, like in question 15 (What is the longest river?), or filtering, like in 92 above. It is especially in these queries that the systems perform poorly on.

5 Conclusion and perspectives

QALD-3, the third edition of the QALD challenge, has attracted a higher number of participants than previous editions, showing that there is a growing interest among researchers to provide end users with an intuitive and easy-to-use access to the huge amount of data present on the Semantic Web—not only by means of classical question answering but also exploiting other paradigms such as faceted search. Although the main focus of the challenge has been on multilinguality, all participating systems worked on English data only. This shows that the multilingual scenario is not yet broadly addressed, although it is starting to attract attention (for a system that used translated QALD questions for evaluation see [1]). Further, the ontology lexicalization task was addressed only by one system during training phase (and one not participating in the challenge, see [15]) but by no participants during test phase. This hints at a slightly different integration of this task into the challenge, e.g. by providing lexica as additional resources for participating question answering system, and by inviting participants to share their own lexical resources.

In future challenges, we want to emphasize further aspects of question answering over linked data, such as the need to deal with a variety of interconnected datasets as well as hybrid sources of information (structured RDF data and unstructured text), while keeping the core task of multilingual question answering. Since the MusicBrainz dataset provided in all three QALD challenges was never used as much as DBpedia, we plan to move to a different domain that can arouse a broader interest. In particular, we think that the biomedical domain has the strong potential to attract new participants and to offer new challenges in the field of question answering over linked data.

References

1. N. Aggarwal, T. Polajnar, and P. Buitelaar. Cross-lingual natural language querying over the web of data. In *Natural Language Processing and Information Systems*, pages 152–163. Springer, 2013.
2. R. Basili and F. M. Zanzotto. Parsing engineering and empirical robustness. *Natural Language Engineering*, 8:2002, 2002.
3. C. Dima. Intui2: A prototype system for question answering over linked data. In *Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF2013. Lecture Notes in Computer Science (to appear)*. Springer, 2013.
4. C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.
5. S. Ferré. squall2sparql: a translator from controlled english to full sparql 1.1. In *Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF2013. Lecture Notes in Computer Science (to appear)*. Springer, 2013.
6. C. Giannone, V. Bellomaria, and R. Basili. A hmm-based approach to question answering against linked data. In *Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF2013. Lecture Notes in Computer Science (to appear)*. Springer, 2013.
7. J. Guyonvarch and S. Ferré. Scalewelis: a query-based faceted search system on top of sparql endpoints. In *Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF2013. Lecture Notes in Computer Science (to appear)*. Springer, 2013.
8. S. He, S. Liu, Y. Chen, G. Zhou, K. Liu, and J. Zhao. Casia@qald-3: A question answering system over linked data. In *Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF2013. Lecture Notes in Computer Science (to appear)*. Springer, 2013.
9. D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.
10. V. Lopez, C. Unger, P. Cimiano, and E. Motta. Evaluation question answering over linked data. *Journal of Web Semantics*, in press.
11. V. Lopez, V. S. Uren, M. Sabou, and E. Motta. Is question answering fit for the semantic web?: A survey. *Semantic Web*, 2(2):125–155, 2011.
12. N. Nakashole, G. Weikum, and F. Suchanek. Patty: a taxonomy of relational patterns with semantic types. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12*, pages 1135–1145, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
13. J. Nivre, J. Hall, and J. Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *In Proc. of LREC-2006*, pages 2216–2219, 2006.
14. C. Pradel, G. Peyet, O. Haemmerlé, and N. Hernandez. Swip at qald-3: results, criticisms and lesson learned. In *Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF2013. Lecture Notes in Computer Science (to appear)*. Springer, 2013.
15. S. Walter, C. Unger, and P. Cimiano. A corpus-based approach for the induction of ontology lexica. In *Natural Language Processing and Information Systems*, pages 102–113. Springer, 2013.