

Trimming a consistent OWL knowledge base, relying on linguistic evidence

Julien Corman
IRIT - Toulouse
julien.corman@irit.fr

Nathalie Aussenac-Gilles
IRIT, CNRS - Toulouse
ausсенac@irit.fr

Laure Vieu
IRIT, CNRS - Toulouse
LOA - Trento
vieu@irit.fr

Abstract

Intuitively absurd but logically consistent sets of statements are common in publicly available OWL datasets. This article proposes an original and fully automated method to point at erroneous axioms in a consistent OWL knowledge base, by weakening it in order to improve its compliance with linguistic evidence gathered from natural language texts. A score for evaluating the compliance of subbases of the input knowledge base is proposed, as well as a trimming algorithm to discard potentially erroneous axioms. The whole approach is evaluated on two real datasets, with automatically retrieved web pages as a linguistic input.

Introduction

As they grow in size, knowledge bases (KBs) tend to contain statements which may make sense individually but, when taken together, violate common sense intuitions. As an illustration, consider the following set Ω of Description Logics (DL) formulas, issued from DBpedia (Mendes et al., 2012) :

- Ex 1.** $\Omega = \{$
- (1) `owningCompany(Smithsonian Networks, Smithsonian Institution),`
 - (2) `doctoralAdvisor(Thaddeus S.C. Lowe, Smithsonian Institution),`
 - (3) `doctoralAdvisor(Nick Katz, Bernard Dwork),`
 - (4) $\top \sqsubseteq \forall \text{doctoralAdvisor} . \text{Person},$
 - (5) $\top \sqsubseteq \forall \text{owningCompany} . \text{Company} \}$

From (1), (2), (4) and (5), the individual *Smithsonian Institution* must be an instance of both `Company` and `Person`, which may seem counterintuitive, and indeed does not correspond to the overall understanding of these two concepts within DBpedia. This kind of issue is common among OWL datasets, which should not be a surprise. The most conventional way of spotting this kind of errors in OWL is by checking consistency or coherence¹ of the input KB, but negation (or cardinality restriction) is underused in practice. As an illustration, according to the LODstats survey tool (Auer et al., 2012), which provides statistics about a sample of the Linked Open Data (LOD) cloud, the two most standard OWL constructs expressing negation, namely `owl:disjointWith` and `owl:complementOf`, have been observed 333 times and twice respectively, against more than 89 000 occurrences for `owl:subclassOf`.

Let us assume that Ω is part of a larger KB K , for instance a subset of DBpedia extracted for a specific application, or a set of OWL statements aggregated from multiple sources. Assume also that there are several other instances of `Person` and `Company` according to K and, to keep the example simple, that *Smithsonian Institution*, *Bernard Dwork*, `doctoralAdvisor`, and `owningCompany` do not appear in $K \setminus \Omega$. If most instances of `Person` and `Company` according to K are respectively human beings and companies, one can expect the term “the Smithsonian Institution” to appear with linguistic contexts which tend to characterize terms denoting other instances of `Company` according to K (e.g.

¹in the DL sense, i.e. the satisfiability of all atomic concepts

the context “X was established”), but not other instances of `Person` (like “X was born in”). Similarly, “Bernard Dwork” should appear with contexts which are characteristic of terms denoting other instances of `Person` according to K . In other words, by checking the overall compliance of K with some linguistic input, it should be possible to identify some undesirable (`Person(Smithsonian Institution)`) and desirable (`Company(Smithsonian Institution)`, `Person(Bernard Dwork)`) consequences of it.

The next problem consists in determining how K can be weakened in order to discard the former, but keep the latter. Even if one focuses here (for readability) on weakening Ω only, there are several options available. The view adopted here, which is also the most common in the knowledge base debugging literature, is that some axiom(s) of Ω should be discarded, but none of them unnecessarily. Then the only solution in this example consists in discarding (2).

The article investigates the applicability of such a trimming mechanism to moderately large input KBs (up to a few thousand statements), using automatically gathered web pages or snippets as linguistic input. To our knowledge, this is the first attempt to use linguistic evidence in order to automatically weaken an existing KB instead of extending it.

Section 1 reviews existing works in two closely related fields, KB extraction from texts and KB debugging, whereas section 2 introduces some conventions. Section 3 defines a score which evaluates the compliance with the linguistic data of any subbase of the input KB. Section 4 proposes an algorithm to trim the input KB based on this score. Section 5 evaluates the approach with two datasets.

1 State of the art

Knowledge base extraction from texts, or ontology learning (Cimiano, 2006; Buitelaar et al., 2005) aims at automatically building or enriching a knowledge base out of linguistic evidence. The work presented here borrows from a subtask named ontology population (which itself borrows from named entity classification), but only when the individuals and concepts of interest are known in advance (Cimiano and Völker, 2005; Tanev and Magnini, 2008; Giuliano and Gliozzo, 2008), which is a non-standard case, whereas ontology population generally considers retrieving new individuals likely to instantiate a given set of concepts. The objective differs also fundamentally from the one pursued in knowledge base extraction, in that the desired output of the process is a weaker KB from which potentially faulty statements have been discarded, not a stronger one. In that sense, this work pertains to knowledge base debugging, for which different tools or algorithms have been devised in the recent years, performing for instance a series of syntactic verifications (Poveda-Villalón et al., 2012), or submitting models (Ferré and Rudolph, 2012; Benevides et al., 2010) or consequences (Pammer, 2010) of the input KB to the user.

In a more automated fashion, diagnosis for Description Logics (Ribeiro and Wassermann, 2009; Schlobach, 2005; Friedrich and Shchekotykhin, 2005; Kalyanpur et al., 2006; Qi et al., 2008) deals with automated weakening of an input KB. An important difference though between this work and approaches based on diagnosis is that the latter assume that the input KB is inconsistent (or incoherent in the DL sense), or at least that some undesired consequences of K have been previously identified. Another drawback of KB diagnosis without an external source of knowledge (like the linguistic input used here) is the sheer number of candidate subbases, as experienced by (Schlobach, 2005).

2 Conventions

The reader is assumed familiar with the syntax and standard model-theoretic semantic of Description Logics (Baader, 2003). A *knowledge base* (KB) Γ is just a finite set of DL formulas. Following the usage in the DL community, the term *axiom* designates a formula $\phi \in \Gamma$, whereas a *consequence* ψ is a formula verified by all models of Γ . A KB is said consistent iff it admits a model. A DL *atomic concept* designates a unary predicate without logical connective or quantifier, like `Company` or `Person`, as opposed to a *complex concept*, like `\exists doctoralAdvisor. \neg Person`. A DL *role* is a binary predicate.

The method introduced here can in theory be applied to any KB in a DL with available reasoning facilities, in particular the DLs underlying the OWL DL and OWL 2 recommendations, but it is arguably

better-suited for KBs in less expressive DLs, which also constitute the vast majority of the data available on the LOD cloud. Another requirement is the presence of linguistic terms denoting named individuals of the input KB, prototypically given by their OWL labels.

3 Compliance of a set of statements with a linguistic input

This section defines a score which reflects the compliance of a set of statements Γ with a linguistic corpus, and will be used in the next section to identify potentially faulty axioms in an input KB K . More precisely, what this score evaluates is the compliance with the linguistic input of the set $\sigma(\Gamma)$ of all consequences of Γ of the form $A(e)$ or $\neg A(e)$, where A and e are respectively a DL atomic concept and a DL individual, and such that there is at least one (other) instance e' of A according to Γ . For an atomic DL concept A , let $\sigma_A(\Gamma) \subseteq \sigma(\Gamma)$ denote the consequences of Γ of the form $A(e)$. Roughly speaking, for each $\psi = A(e)$ or $\psi = \neg A(e)$ in $\sigma(\Gamma)$, the method exploits linguistic contexts which, according to Γ , are characteristic of instances of A , based on $\sigma_A(\Gamma) \setminus \{e\}$, yielding a score $sc_\Gamma(\psi)$ for ψ . This score reflects how much the linguistic behavior of e resembles or deviates from the linguistic behavior of (other) instances of A . For instance, in example 1, let $\psi = \text{Person}(\text{Smithsonian Institution})$, and assume that Ω is part of a larger (consistent) KB K . Then $\psi \in \sigma_{\text{Person}}(K) \subseteq \sigma(K)$, and the score $sc_K(\psi)$ of ψ is determined by the linguistic contexts shared by the individual *Smithsonian Institution* and other instances of *Person* according to K .

A first important observation is that no assumption is made regarding the veracity of a consequence like $\psi = \text{Person}(\text{Smithsonian Institution})$. This would require some external knowledge about reality, which is beyond the scope of this work. The only source of knowledge is the input KB itself, paired with the linguistic input. For instance, the concept *Person* in K may encompass juridical persons, in which case one should expect the score $sc_K(\psi)$ to be high.

Another remark is that the linguistic term “person” does not play any function here. The label of the atomic concept *Person* could actually be “B.27”, with no incidence on $sc_K(\psi)$. This contrasts with a relatively widespread approach in the knowledge base extraction literature, which consists in looking for (possibly learned) cooccurrence patterns (Hearst, 1992) between terms denoting instances and classes. For instance, if X denotes an individual and Y a concept, then the linguistic patterns “ X is a Y ” or “ X and other Y s” tend to indicate that X is indeed an instance of Y . There are at least two reasons to prefer similarity of contexts to cooccurrences patterns for the precise task addressed here. First, when the instances and classes of interest are known in advance (which is the case here, but usually not for knowledge base extraction), similarity of contexts has been empirically shown to outperform cooccurrence patterns (Tanev and Magnini, 2008). This may be explained by the fact that retrieving a sufficient number of cooccurrences of a given pair of terms (e.g. “Thaddeus S.C. Lowe” and “person”) is not always possible, whereas retrieving simple occurrences of a term (e.g. “Thaddeus S.C. Lowe”) is obviously easier. Cooccurrences can also be harder to retrieve for more abstract concepts. For instance, “Virgin Holydays is [...] company”, or “Virgin Holydays and other companies” both sound plausible, but “Virgin Holydays is [...] organization” or “Virgin Holydays and other organizations” is less likely to be found. The second reason is that terms denoting classes (prototypically common nouns) are arguably more ambiguous than terms denoting individuals (prototypically named entities). This does not completely solve the ambiguity issue though, in particular for homonyms (for instance, the term “JKF” may designate either a politician or an airport), and additional precautions may be taken for these, like the ones described in section 5.

Here is now a concrete proposition for the computation of $sc_\Gamma(\psi)$, given a consistent KB Γ and a consequence $\psi \in \sigma(\Gamma)$. The corpus is constituted of either web pages or snippets retrieved with a search engine, using as queries named entities which denote individuals appearing in Γ , generally given as OWL labels. A linguistic context in this setting is just an n -gram ($2 \leq n \leq 5$) immediately following or preceding a term of interest, and without punctuation mark. Borrowing from (Giuliano and Gliozzo, 2008), the observed frequencies of an n -gram s are adjusted based on the self information $self(s)$, given

by $\text{self}(s) = -\log p(s)$, the probability $p(s)$ being estimated with the Google Web 1T 5-gram corpus²³ (this intuitively penalizes n-grams with a high Google n-gram frequency).

Let Cont be the set of all contexts observed with individuals of Γ , and, if y is an individual, let $\mathbf{y} \in \mathbb{R}^{|\text{Cont}|}$ denote the vector of frequencies of contexts observed with y . The confidence score $\text{sc}_\Gamma(\psi)$ is given either by :

Definition 3.1. $\text{sc}_\Gamma(A(e)) = p(\Gamma \models A(x) \mid \mathbf{x} = \mathbf{e})$

Definition 3.2. $\text{sc}_\Gamma(\neg A(e)) = 1 - p(\Gamma \models A(x) \mid \mathbf{x} = \mathbf{e})$

Intuitively, $p(\Gamma \models A(x) \mid \mathbf{x} = \mathbf{e})$ denotes the probability, for a random individual x with the same context frequency vector as e , that $\Gamma \models A(x)$. Applying Bayes' rule, it is equivalent to :

$$p(\mathbf{x} = \mathbf{e} \mid \Gamma \models A(x)) \cdot \frac{p(\Gamma \models A(x))}{p(\mathbf{x} = \mathbf{e})}$$

Let $\text{inst}(\Gamma, C)$ denote the instances of concept C according to Γ , and $\sum \mathbf{y}$ the cumulated values of vector \mathbf{y} . Then $p(\Gamma \models A(x))$ can be estimated by $\frac{\sum_{x \in \text{inst}(\Gamma, A)} \sum \mathbf{x}}{\sum_{x' \in \text{inst}(\Gamma, \top)} \sum \mathbf{x}'}$.

Estimating $p(\mathbf{x} = \mathbf{e} \mid \Gamma \models A(x))$ is slightly more complex. Let $\cos(\mathbf{y}, \mathbf{y}')$ designate the cosine similarity between vectors \mathbf{y} and \mathbf{y}' . If $X_{\mathbf{y}}^{\mathbf{y}'}$ is a random variable for $\cos(\mathbf{y}, \mathbf{y}')$, then $p(X_{\mathbf{y}}^{\mathbf{y}'} \leq \cos(\mathbf{y}, \mathbf{y}'))$ indicates how unexpectedly similar the observed linguistic behaviors of individuals y and y' are.

Then $p(\mathbf{x} = \mathbf{e} \mid \Gamma \models A(x))$ can be estimated by $\sum_{e' \in \text{inst}(\Gamma, A) \setminus \{e\}} \frac{p(X_{\mathbf{e}}^{\mathbf{e}'} \leq \cos(\mathbf{e}, \mathbf{e}'))}{|\text{inst}(\Gamma, A) \setminus \{e\}|}$.

And similarly, $p(\mathbf{x} = \mathbf{e})$ can be estimated by $\sum_{e' \in \text{inst}(\Gamma, \top) \setminus \{e\}} \frac{p(X_{\mathbf{e}}^{\mathbf{e}'} \leq \cos(\mathbf{e}, \mathbf{e}'))}{|\text{inst}(\Gamma, \top) \setminus \{e\}|}$.

For the experiments described in section 5, $p(X_{\mathbf{e}}^{\mathbf{e}'} \leq \cos(\mathbf{e}, \mathbf{e}'))$ was computed by assuming a normal distribution of $X_{\mathbf{e}}^{\mathbf{e}'}$, whose parameters were estimated by maximum likelihood out of all pairwise cosine distances between individuals appearing in $\sigma(\Gamma)$.

The linguistic compliance score $\text{sc}(\Gamma)$ for a set Γ of DL statements can then be defined as the mean of the scores of all $\psi \in \sigma(\Gamma)$:

Definition 3.3. $\text{sc}(\Gamma) = \sum_{\psi \in \sigma(\Gamma)} \frac{\text{sc}_\Gamma(\psi)}{|\sigma(\Gamma)|}$

4 Trimming a KB using linguistic compliance

This section shows how the compliance score which has just been defined can be used to refine an input KB K , in order to rule out potentially faulty axioms. Ideally, one would like to identify the subbases of K which are maximal wrt to set inclusion among the ones having a maximal compliance score. But in practice, even if this does not affect the worst-case complexity of the whole trimming process (dominated by the computation of $\sigma(K)$), the size of the search space (2^K) makes the identification of the optimal subbase(s) of K hardly feasible for realistically sized datasets. This is why the heuristic described by algorithm 1 was used for the experiments of section 5.

Intuitively, it consists in incrementally discarding the axiom which penalizes the most the linguistic compliance of the current base Γ , initialized with K . In other words, the selected axiom is the one which, if discarded, yields the immediate subbase with maximal compliance score. The output is a list containing successively discarded axioms, the first discarded ones being considered as the least reliable. The number n of axioms to discard is chosen as a parameter. Two limit cases are ignored for the sake of readability. First, it is assumed that $|\arg\max_{\phi \in \Gamma} \text{sc}(\Gamma \setminus \{\phi\})| = 1$, i.e. that there is a single least reliable

²<https://catalog.ldc.upenn.edu/LDC2006T13>

³Thanks to the linguistic department of the Erlangen-Nürnberg University for allowing us to query this corpus.

axiom at each iteration, which is usually the case if n is relatively small. Otherwise, the procedure can be adapted in a straightforward manner to return a set of lists of axioms, instead of a single list. The other (unlikely) limit case occurs when $\max_{\phi \in \Gamma} \text{sc}(\Gamma \setminus \{\phi\}) \leq \text{sc}(\Gamma)$, i.e. when no immediate subbase of Γ has a better compliance score than Γ . If this happens, the procedure should simply be interrupted, returning only the axioms discarded thus far.

Algorithm 1 Trimming heuristic

```

1: OutputList  $\leftarrow$  EmptyList ;
2:  $\Gamma \leftarrow K$  ;
3: while  $|\Gamma| > |K| - n$  do
4:    $Ax \leftarrow \underset{\phi \in \Gamma}{\text{argmax}} \text{sc}(\Gamma \setminus \{\phi\})$  ;
5:   append(OutputList,  $Ax$ ) ;
6:    $\Gamma \leftarrow \Gamma \setminus \{\phi\}$  ;
7: end while

```

This procedure is only a heuristic, in that there is no guarantee that the complement in K of the n discarded axioms has an optimal compliance score among subbases of K (not even among subbases of K with cardinality $\geq |K| - n$).

An alternative approach was also experimented, which captures a slightly different intuition, namely that the loss of consequences of $\sigma(\Gamma)$ with low scores should be prioritized when selecting the axiom Ax to discard, line 4. Let $f(\phi) \in \mathbb{R}^*$ be the ordered list of all $\text{sc}_\Gamma(\psi)$ for $\psi \in \sigma(\Gamma \setminus \{\phi\})$. Then if \preceq_l is the standard lexicographic order over \mathbb{R}^* , the axiom ϕ is discarded iff $\forall \phi' \in \Gamma \setminus \{\phi\} : f(\phi) \prec_l f(\phi')$.

5 Evaluation

Both approaches were evaluated with 2 consistent datasets, using a distinct evaluation protocol for each dataset. The first dataset is an automatically retrieved subset of DBpedia, thematically focused on tourism. It counts 5721 logical axioms, and 1095 DL named individuals, with relatively simple formulas (the least expressive underlying DL is $\mathcal{AL}^{(D)}$). This is an example of a lightweight KB, with a large predominance of ABox axioms (5336 over 5721). Additionally, it is a fragment of a large dataset (DBpedia) mainly built out of semi-structured data (Wikipedia infoboxes), but also partly issued from a collaborative effort (the DBpedia ontology), and therefore it is likely to contain nonsensical sets of statements. The procedure applied to obtain this KB is described in (Corman et al., 2015). In particular, individuals with potential homonyms (like *JFK*) have been discarded based on the existence of a Wikipedia disambiguation page, of other named individuals sharing their label in DBpedia, or simply if the number of matched web pages for (one of) the label(s) of the individual was too high. The corpus for this dataset was composed of approximately 60000 web pages retrieved with a search engine, using named individual labels as queries. The evaluation consisted in manually verifying whether a discarded axiom ϕ was actually erroneous, i.e. whether the understanding of some element of the signature of ϕ (named individual, atomic concept or role) was incompatible with its overall understanding within K .

The results of this first evaluation are presented in table 1. Columns “5”, “10” and “20” give the number (“val.”) and proportion (“prec.”, for precision) of axioms manually identified as actually erroneous among the 5, 10 and 20 first discarded ones. Column “Ordering” specifies the method applied to select the discarded axiom ϕ at each iteration of the loop in algorithm 1 : “ $\text{sc}(\Gamma \setminus \{\phi\})$ ” if $\phi = \underset{\phi \in \Gamma}{\text{argmax}} \text{sc}(\Gamma \setminus \{\phi\})$, and “ \preceq_l ” if ϕ is obtained with the alternative approach (the lexicographic order) presented in section 4. The values obtained are encouraging, in that one can reasonably expect the proportion of erroneous within the whole KB to be much lower than the precision scores obtained here. A more thorough examination of the linguistic contexts responsible for these good results is still required though.

Ordering	5		10		20	
	val.	prec.	val.	prec.	val.	prec.
$sc(\Gamma \setminus \{\phi\})$	3	0.6	3	0.3	9	0.45
\preceq_l	2	0.4	4	0.4	8	0.4

Table 1: Actually erroneous axioms among the first 5, 10 and 20 discarded ones for the DBpedia subset

Ordering	5		10		20	
	val.	prec.	val.	prec.	val.	prec./rec.
$sc(\Gamma \setminus \{\phi\})$	5	1	6	0.6	9	0.45
\preceq_l	3	0.6	7	0.7	11	0.55

Table 2: Random axioms among the first 5, 10 and 20 discarded ones for the fragment of the fisheries ontology

The second dataset is a small randomly extracted fragment of the fisheries KB built for the NEON project,⁴ which contains 169 logical axioms, involving only 20 named individuals (mostly geographical or administrative entities), with a more complex TBox (the least expressive underlying DL is \mathcal{ST}). This KB is arguably more reliable too, which allowed the experimentation of a more objective form of evaluation. It consists in artificially extending K with randomly generated axioms, before trying to discard them by application of the trimming algorithm. The assumption is that a randomly generated axiom is usually less reliable than a manually crafted one. The axiom generation procedure randomly selects an axiom $\phi \in K$, and yields an axiom ϕ' with the same syntactic structure, but in which all individuals, atomic concepts and roles have been randomly replaced by individuals, atomic concepts and roles appearing in K . For instance, if $\phi = A \sqsubseteq \forall r. \neg B$, then $\phi' = C \sqsubseteq \forall s. \neg D$, with C and D (resp. s) randomly chosen among atomic concepts (resp. roles) of the signature of K . Additionally, if $(\phi'_1, \dots, \phi'_n)$ designate the random axioms successively added to K , it was required for each ϕ'_i that $K \cup \{\phi'_1, \dots, \phi'_i\}$ was consistent, and that there was at least one consequence of the form $A(e)$ or $\neg A(e)$ entailed by $K \cup \{\phi'_1, \dots, \phi'_i\}$ but not by $K \cup \{\phi'_1, \dots, \phi'_{i-1}\}$, and such that e shares at least one n-gram with some other named individual of the signature of K . 20 axioms in total were added to K . The corpus consisted of approximately 4500 web pages, retrieved in the same way as for the first dataset. Results are presented in table 2. This time, the values are the number and proportion of randomly generated axioms among the first 5, 10 and 20 discarded ones. Because the numbers of generated and trimmed axioms are identical (20), column “prec./rec.” estimates both precision and recall. Precision was high for the first discarded axioms when $sc(\Gamma \setminus \{\phi\})$ was used to order immediate subbases of Γ . But in both cases, the number of randomly generated axioms among the 20 first discarded ones was not significant.

Conclusion

This article proposes an original approach to identify potentially faulty axioms within a (lightweight) OWL knowledge base, trimming it in order to improve its compliance with some automatically gathered linguistic evidence. A score is defined to evaluate the compliance with the linguistic data of subbases of the input KB, exploiting contexts shared by individuals which, according to a subbase, are instances of the same atomic DL concept. An incremental trimming strategy based on this score is then proposed and evaluated.

⁴<http://www.neon-project.org/nw/Ontologies>

References

- Auer, S., J. Demter, M. Martin, and J. Lehmann (2012). LODStats—an extensible framework for high-performance dataset analytics. In *Knowledge Engineering and Knowledge Management*. Springer.
- Baader, F. (2003). *The description logic handbook: theory, implementation, and applications*. Cambridge university press.
- Benevides, A., G. Guizzardi, B. Braga, and J. Almeida (2010). Validating modal aspects of OntoUML conceptual models using automatically generated visual world structures. *Journal of Universal Computer Science* 16(20).
- Buitelaar, P., P. Cimiano, and B. Magnini (2005). *Ontology Learning from Text: Methods, Evaluation And Applications*. IOS Press.
- Cimiano, P. (2006). *Ontology learning and population from text: algorithms, evaluation and applications*. Springer.
- Cimiano, P. and J. Völker (2005). Towards large-scale, open-domain and ontology-based named entity classification. In *RANLP proceedings*.
- Corman, J., L. Vieu, and N. Aussenac-Gilles (2015). Ontological Analysis For Description Logics Knowledge Base Debugging. In *CommonSense proceedings*.
- Ferré, S. and S. Rudolph (2012). Advocatus Diaboli—Exploratory Enrichment of Ontologies with Negative Constraints. *EKAW proceedings*.
- Friedrich, G. and K. Shchekotykhin (2005). A general diagnosis method for ontologies. In *ISWC proceedings*.
- Giuliano, C. and A. Gliozzo (2008). Instance-based ontology population exploiting named-entity substitution. In *COLING proceedings*.
- Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *COLING proceedings*.
- Kalyanpur, A., B. Parsia, E. Sirin, and B. Cuenca-Grau (2006). Repairing unsatisfiable concepts in OWL ontologies. In *ESWC proceedings*.
- Mendes, P. N., M. Jakob, and C. Bizer (2012). DBpedia: A Multilingual Cross-domain Knowledge Base. In *LREC proceedings*.
- Pammer, V. (2010). *Automatic Support for Ontology Evaluation*. Ph. D. thesis, Graz University of Technology.
- Poveda-Villalón, M., M. C. Suárez-Figueroa, and A. Gómez-Pérez (2012). Did you validate your ontology? OOPS! In *ESWC proceedings*.
- Qi, G., P. Haase, Z. Huang, Q. Ji, J. Z. Pan, and J. Völker (2008). A kernel revision operator for terminologies - algorithms and evaluation. In *ISWC proceedings*.
- Ribeiro, M. M. and R. Wassermann (2009). Base revision for ontology debugging. *Journal of Logic and Computation* 19(5).
- Schlobach, S. (2005). Diagnosing terminologies. In *AAAI proceedings*.
- Tanev, H. and B. Magnini (2008). Weakly supervised approaches for ontology population. In *conference on Ontology Learning and Population proceedings*.