Detecting Quality Problems in Semantic Metadata without the Presence of a Gold Standard

Yuangui Lei, Andriy Nikolov, Victoria Uren, and Enrico Motta

Knowledge Media Institute (KMi), The Open University, Milton Keynes,
{y.lei, a. nikolov, v.s.uren, e.motta}@open.ac.uk

Abstract. Detecting quality problems in semantic metadata is crucial for ensuring a high quality semantic web. Current approaches are primarily focused on the algorithms used in semantic metadata generation rather than on the data themselves. They typically require the presence of a gold standard and are not suitable for assessing the quality of semantic metadata. This paper proposes a novel approach, which exploits a range of knowledge sources including both domain and background knowledge to support semantic metadata evaluation without the need of a gold standard. We have conducted a set of preliminary experiments, which show promising results.

1 Introduction

Because poor quality data can destroy the effectiveness of semantic web technology by hampering applications from producing accurate results, detecting quality problems in semantic metadata is crucial for ensuring a high quality semantic web. State-of-art approaches are primarily focused on the assessment of algorithms used in data generation rather than on the data themselves. Examples include the GATE evaluation model [3], the learning accuracy (LA) metric model [2], and the balanced distance metric (BDM) model [11].

As pointed out by [5], semantic metadata evaluation differs significantly from metadata generation algorithms. In particular, the gold standard based approaches that are often used in algorithm evaluation are not suitable for two main reasons. First, it is simply not feasible to obtain gold standards from all the data sources involved, especially, when the semantic metadata are large scale. Second, the gold standard based approaches are not applicable to *dynamic* evaluation, where the process needs to take place on the fly without prior knowledge about data sources.

The approach proposed in this work addresses this issue by exploiting a range of available knowledge sources. In particular, two types of knowledge source are used. One is the knowledge sources that are available in the problem domain, including ontologies. The other type is background knowledge, which includes knowledge sources that are available globally for all applications, e.g., knowledge

sources published on the (Semantic) Web. A set of preliminary experiments have been conducted, which indicate promising results.

The rest of the paper is organized as follows. We begin in Section 2 by describing the motivation of this work in the context of a use scenario. We then present an overview of the approach in Section 3. Next in Section 4 and Section 5, we describe how to exploit each type of knowledge to support the evaluation task. We then describe the settings and the results of the experiments we carried out in this work in Section 6. Finally, we conclude with the key contributions and future work in Section 7.

2 Motivating Scenario: Ensuring High Quality for Semantic Metadata Acquisition

This work was motivated by our work on building a Semantic Web (SW) portal for KMi that would provide an integrated access to resources about various aspects of the academic life of our lab[1]. The relevant data is spread over several different data sources such as departmental databases, knowledge bases and HTML pages. In particular, KMi has an electronic newsletter[2], which describes events of significance to the KMi members. New entries are kept being added to the archive.

There are two essential activities involved in the portal, including i) extracting named entities (e.g., people, organizations, projects, etc.) from news stories in an automatic manner and ii) verifying the derived data to ensure that only data at high quality proceeds to the semantic metadata repository. Both activities take place *dynamically* on a *continuous* basis whenever new information becomes available. In particular, the involved data source is unknown to the portal prior to the metadata acquisition process. Hence, traditional gold standard based evaluation approaches are not applicable, as pre-constructing gold standards is simply not possible.

Please note that although it is drawn from the context of semantic metadata acquisition, the scenario also applies to generic semantic web applications, where evaluation often needs to be performed in an automated manner in order to filter out poor quality data dynamically whenever intermediate results are produced.

3 An Overview of the Proposed Approach

The goal of the proposed approach is to automatically detect data deficiencies in semantic metadata without having to construct gold standard data sets. It was inspired by our previous work ASDI [9], which employs different types of knowledge sources to verify semantic metadata. We extend this method towards a more powerful mechanism to support the checking of data quality by exploiting more types of knowledge sources and by addressing more types of data deficiencies.

[1] http://semanticweb.kmi.open.ac.uk
[2] http://kmi.open.ac.uk/news

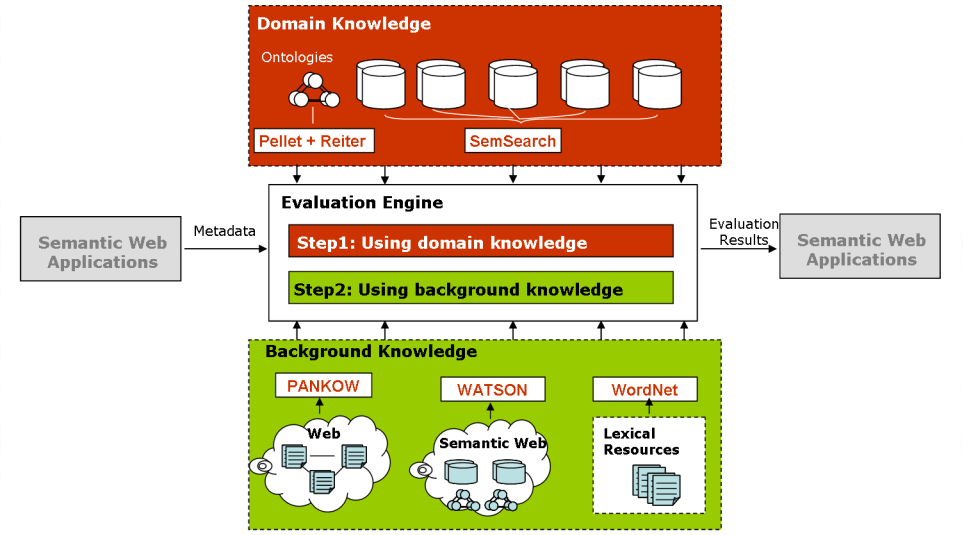

Fig. 1. An Overview of the Proposed Evaluation Approach

Figure 1 shows an overview of the proposed approach. In the following sub sections, we first describe the deficiencies addressed by the proposed approach. We then clarify the knowledge sources used in detecting quality problems.

3.1 Data Deficiencies Addressed

To clarify, we define semantic metadata as RDF triples that describe the meaning of data sources (i.e. semantic annotations) or denote real world objects (e.g., projects and publications) using the specified ontologies. In our previous work, we have developed a quality framework, called SemEval [13], which has identified a set of important data deficiencies that occur in semantic metadata, including:

- **Incomplete annotation**, which defines the situation where the mapping from the objects described in the data source to the instances contained in the semantic metadata set is not exhaustive.
- **Inconsistent annotation**, which denotes the situation where entities are inconsistent with the underlying ontologies. For example, an organization ontology may define that there should be only one director for an organization. The inconsistency problem occurs when there are two directors in the semantic metadata set.
- **Duplicate annotation**, which describes the deficiency in which there is more than one instance referring to the same object. An example situation is that the person *Clara Mancini* is annotated as two different instances, for example *Clara Mancini* and *Clara*.

- **Ambiguous annotation**, which expresses the situation where an instance of the semantic metadata set can be mapped back to more than one real world object. One example would be the instance *John* (of the class *Person*) in the context where there are several people described in the same document who have the name.
- **Inaccurate annotation**, which defines the situation where the object described by the source has been correctly picked up but not accurately described. An extreme scenario in this category is *mis-classification*, where the data object has been successfully picked up and been associated with a wrong class. An example would be the *Organization* instance *Sun Microsystem* marked as a person.
- **Spurious annotation**, which defines the deficiency where there is no object to be mapped back to for an instance. For example, the string *"Today"* annotated as a person.

The proposed approach is designed to address all these data deficiencies except the first one. This is because the approach concentrates especially on the quality status of the semantic annotations that are *already* contained in the given semantic metadata set.

3.2 Knowledge Sources Exploited

As shown in Figure 1, two types of knowledge sources are exploited to support the evaluation task, namely *domain knowledge* and *background knowledge*.

Domain knowledge. Three types of knowledge sources are often available in the problem domain: i) domain ontologies, which model the problem domain and offer rules and constraints for detecting conflicts and inconsistencies contained in the evaluated data set; ii) semantic data repositories, which contain facts of the problem domain that can be looked upon to examine problems like inaccuracy, ambiguity, and inconsistency; and iii) lexical resources, which contain domain specific lexicons that can be used to link the evaluated semantic metadata with specific domain entities. As will be detailed in Section 4, domain knowledge is employed to detect inconsistency, duplicate, ambiguous and inaccurate annotation problems.

Background knowledge. The knowledge sources that fall into this category include: i) online ontologies and data repositories, ii) online textual resources, and iii) general lexical resources. The first two types of knowledge sources are exploited to detect possible deficiencies that might be associated with those entities that are not included in the problem domain (i.e., those entities that do not have matches). General lexical resources, on the other hand, are employed to expand queries when finding matches of the evaluated entity.

Compared to domain knowledge, one characteristic of background knowledge is that it is generic and is available to all applications. Another important feature of the knowledge, especially the first two types of background knowledge, is that they are less trustworthy than domain specific knowledge as the (semantic) web is an open environment where anyone can contribute. Corresponding to the two

types of knowledge sources exploited, the deficiency detection process comprises two major steps, which are described in the following sections.

4 Detecting Data Deficiencies Using Domain Knowledge

The tasks involved in this step are centered around the detection of four types of quality problems that are common to semantic metadata, namely inconsistent, duplicate, ambiguous, and inaccurate problems. The process starts with the detection of inconsistencies that may exist between the evaluated semantic metadata entity with the data contained in the specified semantic data repositories. It then investigates the duplicate problem using the same annotation context. The third task involved is detecting ambiguous and inaccurate problems by querying the available semantic data repositories.

Detecting inconsistencies. Please note that we are only interested in data inconsistencies at the ABox level. Such inconsistencies may be caused by disjointness axioms or the violation of property restrictions. First, disjointness leads to inconsistency when the same individual belongs to two disjoint classes at the same time. For example, the annotation "Ms Windows is a Person" is inconsistent with the statement that defines it as a technology, as the two classes are disjoint with each other. Second, violation of property restrictions (e.g., domain/range restriction, cardinality restriction) also causes inconsistencies. For example, if the ontology defines that there should be only one director in an organization, there is an inconsistency if two people are classified as director.

To achieve the task of inconsistency detection, we employ ontology diagnosis techniques. Each inconsistency is represented by a so-called minimal inconsistent subontology (MISO) [7], which includes all statements and axioms that contribute to the conflict. An OWL-reasoner with explanation capability is able to return a MISO for the first inconsistency found in the data set. The process starts with locating a single inconsistency using the Pellet OWL reasoner [8]. It then discovers all the inconsistencies by using Reiter's hitting set tree algorithm [12], which builds a complete consistent tree by removing each ABox axiom from the MISO one by one. Please see [12] for the detail.

Detecting duplicate problems. This task is achieved by seeking matches of the evaluated entity within the same annotation context, i.e., within the values of the same property of the same instance that contains the evaluated entity. For example, when evaluating the annotation *(story x, mentions-person, enrico)*, the proposed approach examines other person entities mentioned in the same story for detecting the duplicate problem. Domain specific lexicons are used in the process (e.g., the string "OU" stands for "Open University") to address domain specific abbreviations and terms.

Detecting ambiguous and inaccurate problems. This task is fulfilled by querying the available data repositories. When there is more than one match found, the evaluated entity is considered to be ambiguous, as its meaning (i.e., the mapping to real world data objects) is not clear. For example, in the case of evaluating the person entity "John", there is more than one match found in the KMi domain

repository. The meaning of the instance needs to be disambiguated. In the situation where there is an inexact match, the entity is computed as inaccurate. As to the third possibility where there is no match found, the proposed approach turns to background knowledge to carry out further investigation. We used SemSearch [10], a semantic search engine, to query the available data repositories, and a suite of string matching mechanisms to refine the matching result.

5 Checking Entities Using Background Knowledge

There are three possibilities when matches could not be found for the evaluated entity in the problem domain. One is that the entity is correct but not included in the problem domain (e.g., IBM, BBC, and W3C with respect to the KMi domain). The second possibility is *mis-classification*, where the entity is wrongly classified, e.g., "Sun Microsystems" as a "person". The third one is *spurious annotation*, in which the entity is erroneous, e.g., "today" as a "person". Hence, this step focuses on detecting two types of quality problems: mis-classification) and spurious annotation.

The task is achieved by computing possible classifications using knowledge sources published on the (semantic) web. The process begins by querying the semantic web. If satisfactory evidence cannot be derived, the approach then turns to textual resources available on the web (i.e., the general web) for further investigation. If both attempts fail, the system considers the evaluated entity *spurious*.

We used i) *WATSON* [4], a semantic search tool developed in our lab, to seek classifications of the evaluated term from the semantic web; and ii) *PANKOW* [1], a pattern-based term classification tool, to derive possible classifications from the general web. *Detecting mis-classification problems* is achieved by comparing the derived classifications (e.g., *company* and *organization* in the case of evaluating the annotation "Sun Microsystems as person") to the type of the evaluated entity (which is the class *person* in the example) by exploring domain ontologies and general lexicon resources like WordNet [6]. In particular, the disjointness of classes are used to support the detection of the problem. General lexicon resources are also exploited to compute the semantic similarities of the classifications.

6 Experiments

In this work we have carried out three preliminary experiments, which investigate the performance of the proposed approach in the KMi domain. In the following subsections, we first describe the settings and the methods of the experiments. We then discuss the results of the experiments.

6.1 Setup

The experimental data were collected from the previous experiment carried out in ASDI [9], in which we randomly chose 36 news stories from the KMi news

archive[3] and constructed a gold standard annotation collection by asking several KMi researchers to manually mark them up in terms of *person, organization* and *projects*. We used the semantic metadata set that was automatically gathered from the chosen news stories by the named entity recognition tool ESpotter [14] as the data set that needs evaluation. We then experimented with this semantic metadata set using a gold standard based approach and the proposed approach.

In order to get a better idea of the performance of the proposed approach on employing different types of knowledge sources, we conducted three experiments: the first experiment used the constructed gold standard annotation collection; the second one used domain knowledge sources; and the third experiment used both domain knowledge and background knowledge. In particular, for the purpose of minimizing the influences that may be caused by other factors such has human intervention, we developed automatic evaluation mechanisms for both the gold standard based approach and the proposed approach, which use the same matching mechanism. Table 1 shows the results, with each cell presenting the total number of the correspondent data deficiencies (i.e., row) found in the data set with respect to the extracted entity type (or the the sum of all types).

Table 1. The Data Deficiency Detection Results of the Experiments

Deficiency	**People**	**Organizations**	**Projects**	**Total**
Experiment 1: Using the gold standard annotations				
Incomplete annotation	17	16	9	42
Inconsistent	**n/a**(not applicable)			
Duplicate	3	10	0	13
Ambiguous	0	1	0	1
Inaccurate	0	1	0	1
Spurious	8	17	0	25
*Experiment 2: Using **domain knowledge only***				
Incomplete annotation	**n/a**			
Inconsistent	1	0	0	1
Duplicate	3	10	0	13
Ambiguous	0	1	0	1
Inaccurate	1	3	0	4
Spurious	33	45	2	80
*Experiment 3: Using **both domain knowledge and background knowledge***				
Incomplete annotation	**n/a**			
Inconsistent	**5**	**8**	**0**	**13**
Duplicate	3	10	0	13
Ambiguous	0	1	0	1
Inaccurate	1	3	0	4
Spurious	**5**	**8**	**0**	**13**

[3] http://kmi.open.ac.uk/news

6.2 Discussion

Assessing the performance of the proposed approach is difficult, as it largely depends on three factors, including i) whether it is possible to get hold of good data repositories that cover most facts of the problem domain, ii) whether the relevant topics have gained good publicity on the (semantic) web, and iii) whether the background knowledge itself is of good quality and trustworthy. Here we compare the results of the different experiments in the hope of finding some clues of the performance.

Comparing the proposed approach with the gold standard based approach. As shown in the table, the performances on detecting duplicate, ambiguous and inaccurate problems are quite close. This is because that, like gold standard annotations, the KMi domain knowledge repositories cover all the facts (including people, projects, organizations) that are contained in the domain. On the other hand, there are two major differences between the gold standard based approach and the proposed approach.

One major difference is that, in contrast with the gold standard based approach, the proposed approach is able to detect inconsistent annotations but with no support for the incomplete annotation problem. This is because the proposed approach deliberately includes domain ontologies as a type of knowledge sources and does not have the knowledge of full set of annotations of the data source.

Another major difference lies in the detection of the spurious annotation problem. More specifically, there is a big difference between the first experiment and the second one. This is mainly caused by the fact that many entities extracted from the news stories are not included in the domain knowledge (e.g., "IBM", "BBC", "W3C"), and thus are not being to be covered by the second experiment. But they are contained in the manually constructed gold standard.

There is also a significant difference between the first experiment and the third one with respect to the detection of spurious annotations. Further investigation reveals two problems. One is that the gold standard data set is not perfect. Some entities are not included but correctly picked up by the extraction tool. "EU Commissioner Reding as a person" is such an example. The other problem is that background knowledge can sometimes lead to false conclusions. On the one hand, some spurious annotations are computed as correct, due to the difficulties in distinguishing different senses of the same word in different contexts. For example, "international workshop" as an instance of the class *Organization* is computed as correct, whereas the meaning of the word *organization* when associating with the term is quite different from the meaning of the class in the KMi domain ontology. On the other hand, false alarms are sometimes produced due to the lack of publicity of the evaluated entity in the background knowledge. For example, in the KMi SW portal, the person instance *Marco Ramoni* is computed as spurious, as not enough evidence could be gathered to draw a positive conclusion.

Comparing the performance of the approach between using and without using background knowledge. With 12 inconsistencies discovered and 58 spurious prob-

lems cleared among the 80 spurious problems detected in the second experiment, the use of background knowledge has proven to be effective in problem detection in the KMi domain. This is mainly because the relevant entities that are contained in the chosen news stories collection have gained fairly good publicity. "Sun Microsystem", "BBC" and "W3C" are such examples. As such, classifications can be easily drawn from the (Semantic) web to support the deficiency detection task. However, as described above, we have also observed that several false results have been produced by the proposed approach.

In summary, the results of the experiments indicate that the proposed approach works reasonably well for the KMi domain when considering zero human effort is required. In particular, domain knowledge is proven to be useful in detecting those problems that are highly relevant to the problem domain, such as ambiguous and inaccurate annotation problems. The background knowledge, on the other hand, is quite useful for investigating those entities that are outside.

7 Conclusions and Future Work

The key contribution of this paper is the proposed approach, which, in contrast with existing approaches that typically focus on the evaluation of semantic metadata generation algorithms, pays special attention to the quality evaluation of semantic metadata themselves. It addresses the major drawback of current approaches suffered when applying to data evaluation, which is the need for gold standards, by exploiting a range of knowledge sources.

In particular, two types of knowledge source are used. One is the knowledge sources that are available in the problem domain, including domain ontologies, domain specific data repositories and domain lexical resources. They are used to detect quality problems of those semantic metadata that are contained in the problem domain, including data inconsistencies, duplicate, ambiguous and inaccurate problems. The other type is background knowledge, which includes ontologies and data repositories published on the semantic web, online textual resources, and general lexical resources. It is mainly used to detect quality problems that are associated with those data that are not contained in the problem domain, including mis-classification and spurious annotations.

We have conducted three preliminary experiments examining the performance of the proposed approach, with each focusing on the use of different types of knowledge sources. The study shows encouraging results. We are, however, aware of *a number of issues* associated with the proposed approach. For example, real time response is crucial for dynamic evaluation, which takes places at run time. How to speed up the evaluation process is an issue that needs to be investigated in the future. Another important issue is the impact of the trustworthiness of different types of knowledge on the evaluation.

Acknowledgements

This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

References

1. P. Cimiano, S. Handschuh, and S. Staab. Towards the Self-Annotating Web. In *Proceedings of the 13th International World Wide Web Conference*, pages 462 – 471, 2004.
2. P. Cimiano, S. Staab, and J. Tane. Acquisition of Taxonomies from Text: FCA meets NLP. In *Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining*, pages 10 – 17, 2003.
3. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL02)*, 2002.
4. M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. WATSON: A Gateway for the Semantic Web. In *4th European Semantic Web Conference (ESWC'07)*, 2007.
5. K. Dellschaft and S. Staab. Strategies for the Evaluation of Ontology Learning. In P. Buitelaar and P. Cimiano, editors, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. IOS Press, December 2007.
6. C. Fellbaum. *WORDNET: An Electronic Lexical Database*. MIT Press, 1998.
7. P. Haase, F. Harmelen, Z. Huang, H. Stuckenschmidt, and Y. Sure. A framework for handling inconsistency in changing ontologies. In *Proceedings of the International Semantic Web Conference (ISWC2005)*, 2005.
8. A. Kalyanpur, B. Parsia, E. Sirin, and B. Grau. Repairing Unsatisfiable Concepts in OWL Ontologies. In *3rd Eupopean Semantic Web Conference*, pages 170–184, 2006.
9. Y. Lei, M. Sabou, V. Lopez, J. Zhu, V. S. Uren, and E. Motta. An Infrastructure for Acquiring High Quality Semantic Metadata. In *Proceedings of the 3rd European Semantic Web Conference*, 2006.
10. Y. Lei, V. Uren, and E. Motta. SemSearch: A Search Engine for the Semantic Web. In *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management(EKAW2006)*, 2006.
11. D. Maynard, W. Peters, and Y. Li. Metrics for Evaluation of Ontology-based Information Extraction. In *Proceedings of the 4th International Workshop on Evaluation of Ontologies on the Web*, Edinburgh, UK, May 2006.
12. R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987.
13. Enrico Motta Yuangui Lei, Victoria Uren. SemEval: A Framework for Evaluating Semantic Metadata. In *The Fifth International Conference on Knowledge Capture (KCAP 2007)*, 2007.
14. J. Zhu, V. Uren, and E. Motta. ESpotter: Adaptive Named Entity Recognition for Web Browsing. In *Proceedings of the Professional Knowledge Management Conference*, 2004.