Towards Enriching Linked Open Data via Open Information Extraction

Antonis Koukourikos[1,2], Vangelis Karkaletsis[2], George Vouros[1]

[1]University of Piraeus, Department of Digital Systems. 80, Karaoli and Dimitriou Str, Piraeus, 18534
[2]Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos" Agia Paraskevi Attikis, P.O.Box 60228, 15310 Athens, Greece

Abstract. The descriptions of various entities on Linked Data repositories are subject to constant renewals and modifications, with respect to both the descriptions of concepts and relations and entities realizing their instantiations. Thus, the underlying ontologies have to be updated accordingly in order to reflect these changes. This paper presents a system for examining the possibilities of discovering new relations and updating/verifying existing ones for entities described in Linked Data repositories by using Open Information Extraction techniques. These are applied to web content. The process aims to the enrichment of the examined datasets and the expansion of the ontologies with newly-discovered concepts and relations. Towards this target, the paper discusses the intricacies, pitfalls and challenges present.

Keywords: Linked Open Data, Open Information Extraction, Ontology Population, Ontology Enrichment

1 Introduction

The Linked Data initiative aims to provide a set of guidelines and best practices for publishing structured data and associating it with other resources. The Linking Open Data community project [1] aims to publish open data sets as RDF triples and establish RDF links between objects from different data sets.

The steadily increasing amount of the datasets involved in the project[1] and the structured nature of the included information provides a foundation for establishing fast, easy and customizable access to substantial knowledge resources. However, it is important to ensure that the information provided by such repositories is constantly updated and expanded. It is necessary to examine the adequacy of the concepts and relations used for describing any entity as well as assessing the validity of the RDF triples.

A possibly promising approach is to exploit the rich and constantly updated content available in the World Wide Web in order to build methods for tackling the aforementioned issues.

The present paper presents an experimental system towards:

1. discovering new concepts for describing an entity in a LOD repository, thus enriching the underlying ontology of the dataset, by retrieving and analyzing relevant web content;
2. assessing the validity of a property currently present in the LOD repository and discovering new values for that property;
3. adding new instantiations of known properties for a LOD object; i.e. associations of a known type between a known entity and an entity not currently present in the repository.

The paper is structured as follows. First, we provide a brief description of the technologies related to our approach. A presentation of the overall system and the role of each of its components are given in section 3. Preliminary results from experiments with the proposed system are presented in section 4.

[1] http:// stats.lod2.eu/

We conclude with some interesting remarks from the experiments and describe future enhancements and expansions on the system.

2 Related Technologies

An overview of Linked Open Data can be found at [2]. Any open dataset that follows the principles originally set for Linked Data can be said to belong in the LOD universe. The notion of interlinking is very important for the initiative. However, it is still the common case that interlinks between datasets are mainly at the instance level.

Ontology alignment is a key-technology for the purpose of interlinking different datasets, by discovering equivalent and/or semantically similar classes and properties between the ontologies of distinct repositories. Correspondences can then drive further associations between data in distinct repositories. Some examples of recent alignment systems are SAMBO [3], ASMOV [4] and RIMOM [5]. The ontology alignment process may also be instance-based: In this regard, in relation to LOD, approaches aim to connect datasets at the conceptual level using LOD information as it is. The BLOOMS ontology alignment system [6] is such a system. It uses information existing in the Wikipedia hierarchy in order to bootstrap the alignment of two input ontologies

In order to address the expansion of ontologies (either at the conceptual or at the assertional level), we use Open Information Extraction (OIE) methods for retrieving information from the broad spectrum of sources available in the Web. The Open Information Extraction paradigm [7] focuses on obtaining relations between argumentative pairs from unstructured web text, with the best possible accuracy, without introducing distinct training sets or taking into account domain-specific information. Hybrid approaches, which map OIE results to existing ontologies and subsequently use the mapping to improve on the results of the extraction process, are also of particular interest [8, 9].

3 Overall System Architecture

The following figure summarizes the architecture of the experimental system and indicates the individual components.

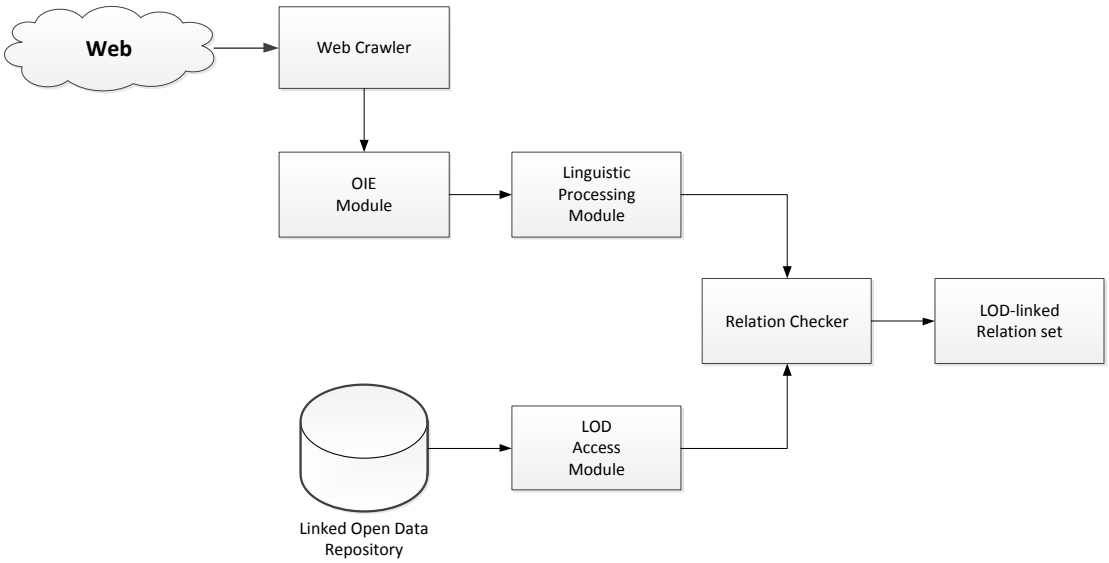

Fig. 1. System architecture

The whole process realized by the system is naturally divided into four stages. The first step is to obtain the raw textual information from the web. The second step is to apply light-weight linguistic manipulation in order to enrich the text with additional information, such as identification of named entities and

co-references, useful for the extraction process. The actual information extraction is employed at the next stage. The extracted information is in the form of a set of triples representing binary relations between entities, is then compared by the relation checker module to the relations obtained by the LOD repositories. A subset of the extracted relations -considered relevant to entities from the repository- is finally produced by the system..

The following section presents the specific technologies that are used for realizing the different components of the system, implementing each of the distinct phases. It also presents specific issues related to setting up the experiments performed.

Linked Open Data Resources: The collection used was the Jamendo dataset, available from DBTune[2], part of the *Linking Open Data in the Semantic Web* project.

The Jamendo dataset contains representations for musicians whose work is available through the Jamendo site, a host of Creative Commons licensed music. The dataset includes 3505 band names and 5786 records. The dataset is linked to the GeoNames[3] dataset for providing the place of origin for an artist.

Collecting the corpus: To activate the crawler, we extracted band names from the Jamendo dataset and performed web searches based on each of these names. From the list of artist names, we selected randomly 200 using a simple random number generator for picking an index on the complete list. The selected names were given as search queries to a module implementing the Bing API. The topmost 50 results for each query were taken into account. The corresponding pages were retrieved and the textual information of the pages was stored. In order to eliminate irrelevant web page content like menus, advertisements and content tables, we implemented a heuristic boilerplate removal module based on the boilerpipe library[4].

Linguistic processing: The first run of the system reveals an important amount of non-resolvable relations due to the use of pronouns or generic phrasing (e.g. “they”, “he”, “the band” etc.). It was deemed necessary to repeat the tests after incorporating a co-reference resolution mechanism. For our first tries, we used the relevant OpenNLP[5] functionalities. Furthermore, the named entity recognition module of the OpenNLP library was re-trained and applied to the corpus.

Information Extraction: The information extraction module that was used was the REVERB system [10], a second generation OIE system which expands the ideas implemented on previous OIE systems by exploiting generic syntactical and lexical constraints in order to reduce the introduction of non-valid relations in the extracted set. As already said, the relations are triples of the form (*Argument, Relation, Argument*), where an argument is an entity related to another one via the indicated relation. For this test, we used the REVERB system with the default settings provided by the creators.

From the set of 10000 pages in the corpus (50 pages for each of 200 artists), REVERB returned a set of 717140 relations. Some of the relations were obviously invalid, since they associated entities with incoherent lexicalizations. We discarded relations for which any of the constituents contained HTML tagging or dynamic elements (JavaScript snippets). Further examination of the relation set will probably reveal additional heuristics for rejecting relations. After this heuristic-based rejection process a set of 506420 relations was considered for further processing.

Association with the LOD dataset: The next step of the experiment was the association of the results from the information extraction with the data available from the Jamendo dataset. The results were classified to the following generic classes:

- Relations on which both arguments were found in the LOD dataset
- Relations on which a single argument was found in the LOD dataset
- Relations that did not associate objects from the LOD dataset

Relations belonging to the first class can either provide an alternate phrasing for a known relation, or introduce a different relation between the known entities (i.e. entities already in the dataset). We used WordNet[6] to examine if the lexicalization of the relation in the LOD repository and the phrase found from the extraction system shared one or more senses. This is done by simple string matching, after trivial changes like removing underscores, punctuation and capitalization. If that is the case, the relation is considered already known. Otherwise, the relation is marked as a possibly new one.

---

[2] http:// dbtune.org/
[3] http:// www.geonames.org/
[4] http://code.google.com/p/boilerpipe
[5] http:// opennlp.sourceforge.net/projects.html
[6] http:// wordnet.princeton.edu/

Each instance-triple of the second class is associating a known entity with an unknown entity, a fact that could also lead to the introduction of an additional concept for describing the unknown entity and of new relations for describing how entities are related.

Instances of the third class were not considered relevant to the purposes of the experiment.

4 Results

The majority of the relations obtained by the information extraction process were not directly relevant to the entities in the Jamendo dataset. 290,714 extracted relations do not concern any entity in the LOD repository (artist name, record name or track name), while in 17,219 of the extracted relations both of their arguments are known entities. The rest of the relations (198,487) indicate a relation with only one known entity, which maybe in the first or in the second argument in the extracted relation triple.

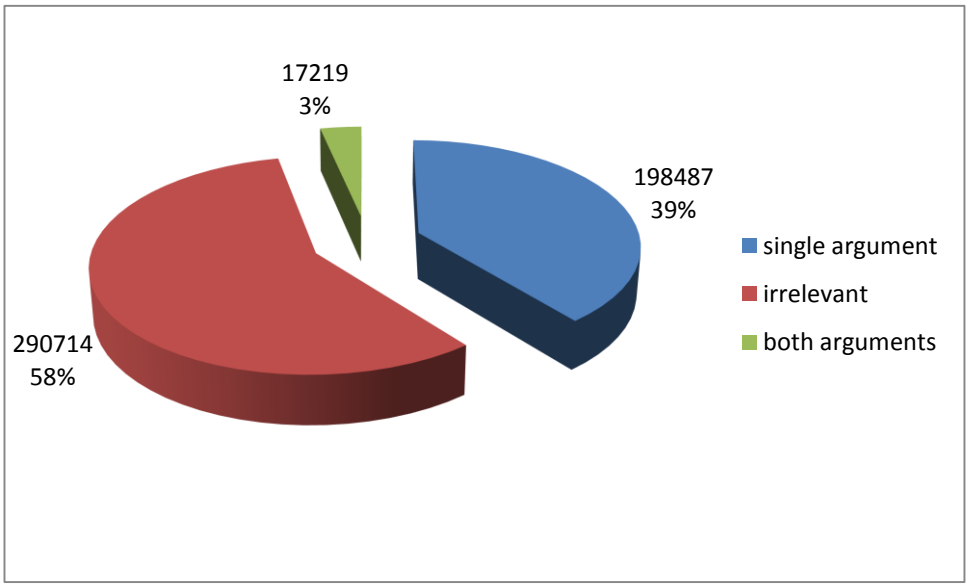

Fig. 2. Distribution of the relations to the classes defined for the experiment

An examination of the test corpus indicated that most irrelevant relations were related to the author of the article or provided general remarks, peripheral to the domain considered; For example generic opinions of the writer, comments on the music industry, social associations of the artist etc.

Some of the entities' properties in the Jamendo RDF dataset have been also retrieved by the OIE. Associations between artist names and their works were the most frequent relations. Such relations are lexicalized in a straightforward manner that is directly resolved by WordNet (e.g. "maker", "creator") and at a context-specific phrasing ("is responsible for", "delivered"). By a brief manual examination of the web corpus we observed that certain information included in the repository was present in the page but it was not provided in a free-formed linguistic style, i.e. in the core text of an article. Rather, the creations of an artist were provided in structured lists or tables, as were the cities on which the musician will perform concerts. It is thus important to consider extraction techniques not based on sentential analysis but also to specific presentation styles, using structured elements (tables or lists), also in combination with information in the context.

The extraction process identified entities which are not included in the Jamendo dataset. These entities may also lead to the inclusion of new concepts/relations in the corresponding ontologies. For instance, the most prevalent was the relation of membership to a band, where a named entity (the musician) was declared as a member, either directly, or by giving his specific role. Another important relation refers to the future releases of records by the band, an element that could also be included in an updated dataset.

With respect to the interlinking with the GeoNames database, it should be mentioned that a significant amount of entities were associated with different geographical names. Such cases were related mostly to tours/concerts and location changes from the artists.

From the NER-enhanced OIE process, we can construct a relation graph for a named entity, with the edges denoting the relations in which the entity participates. This results to a graph of related entities. The dataset used for the experiments limits the generalizing possibilities for the graph, as the information included in the examined web sources do not emphasize on aspects indirectly associated with the music domain. For example, we are not able to find the population or other known inhabitants of a city by only using the relations extracted by the system.

Regarding the available information, it is important to compare the produced relation graphs with corresponding RDF graphs derived from the LOD cloud in order to:

- Deduce the validity of the extracted information
- Disambiguate named entities based on the LOD information and produce distinct graphs for different entities of the same name

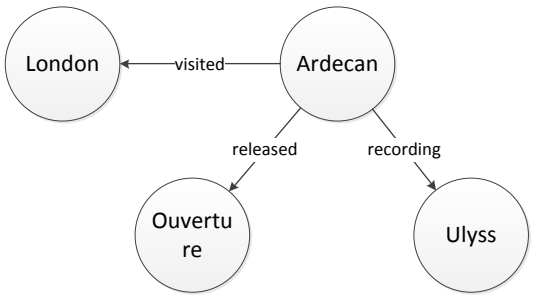

Fig. 3. Relations extracted from the OIE system for an artist

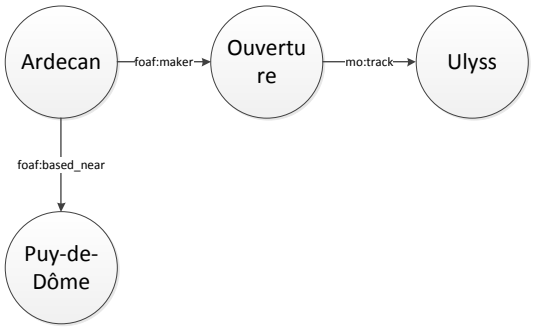

Fig. 4. Part of the RDF description for the same artist

Figures 3 and 4 present information extracted and related to an artist along with a part of the RDF graph for the same artist existing in the Jamendo dataset (The nodes in the RDF graph denote the lexical description of the entities in the RDF triples, for convenience).

While the extraction system discovered an association between the artist and the track, i.e. the relation (Ardecan, recording, Ulyss), the corresponding node for the artist is not correctly associated with the album record (Ouverture) to which the extracted track belongs. From the simple linguistic analysis that we have applied so far, it is not possible to distinguish the role of “Ouverture” and “Ulyss” with respect to the ontology behind the Jamendo dataset.

In addition, the association between the artist and a geographical location that is stated by the “visited” relation is irrelevant to the LOD schema. We expect that this issue will be easily handled with simple some simple linguistic analysis. For example, there is no sense in WordNet that includes both “based” and “visited”, so a cross-check with the WordNet database will reduce the possibility of such relations being similar or equivalent.

5 Conclusions and Future Work

The purpose of our experiments is to examine the possible correlation of linked open data, open information extraction and ontology evolvement.

At this stage, we focused on the possibilities for expanding domain-specific ontologies by using unstructured open-world textual data and associating it with, possibly incomplete, specifications. The initial experiments indicated the presence of multi-faceted information for the entities described in the LOD repositories. The concepts commonly associated with an entity are broader than the ones currently used for its description in the repository, thus the underlying ontology could be expanded in order to include the additional concepts. Furthermore, using linguistic techniques and given an adequate set of external ontologies, we could associate a newly discovered relation with a property in an existing ontology. However this can be arbitrarily intricate. Ontology alignment methods can play a major role towards these goals.

The information extraction process itself should be augmented with techniques that take into account non-sentential web content, as it seems to provide a wealth of information that leads to valid relations.

Our immediate next step is to further analyze the presented results and apply statistical measures in order to deduce whether an extracted relation is actually relevant to the domain and a specific entity, based on the number of occurrences, the association with multiple entities etc. We also aim to refine the linguistic methods applied to the corpus in order to refine the results of the information extraction. These steps shall provide the basis for expressing a methodology for determining graph similarity measures between the relation graph and the LOD-derived RDF graph (or a subset of the latter).

In the long-term, it is important to focus on the relation of the major technologies involved in the system: Open Information Extraction, Linked Open Data and Ontology Enrichment. A gradually improving ontology could be used to assist the information extraction module in order to eliminate irrelevant relations with respect to the domain of the LOD repository. The data in such a repository could also be updated in accordance to the results of the information extraction. Our goal is thus to combine these ideas in a constantly updated system, where the results of the components in a specific run are exploited by subsequent runs in order to increase the efficiency and accuracy of the other components in future executions of the process.

6 Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement n° 288513 (*NOMAD - Policy Formulation through non moderated crowdsourcing*).

7 References

1. http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData. The W3C SWEO community projects home page.
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data – the story so far. International Journal On Semantic Web and Information Systems 5(3), 1–22 (2009)
3. Lambrix, P., Tan, H.: SAMBO – a system for aligning and merging biomedical ontologies. Journal of Web Semantics, vol. 4(1), 196–206, (2006)
4. Jean-Mary, Y. R., Shironoshita, E. P., Kabuka, M. R.: Ontology matching with semantic verification. Journal of Web Semantics, vol. 7(3), 235–251, (2009)
5. Li, J., Tang, J., Li, Y., Luo, Q.: Rimom: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering, vol. 21(8), 1218–1232, (2009)
6. Jain P., Hitzler P., Sheth A.P., Verma K., Yeh P.Z.: Ontology Alignment for Linked Open Data. In proc. of the 9th International Semantic Web Conference (ISWC2010), (2010)
7. Banko M., Cafarella M.J., Soderland S., Broadhead M., Etzioni O.: Open information extraction from the web. International Joint Conference on Artificial Intelligence, (2007)
8. Wu, F., Weld D.S.: Open information extraction using Wikipedia. In proc. of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, 118–127, Morristown, NJ, USA, (2010)
9. Soderland S., Roof B., Qin B., Xu S., Mausam, Etzioni O.: Adapting open information extraction to domain-specific relations. AI Magazine, 31(3), 93–102, (2010)
10. Etzioni O., Fader A., Christensen J., Soderland S, Mausam: Open Information Extraction: the Second Generation. International Joint Conference on Artificial Intelligence, (2011)