Interlinking Cross-Lingual RDF Data Sets

Tatiana Lesnikova

INRIA & LIG, Grenoble, France
{tatiana.lesnikova}@inria.fr
http://exmo.inrialpes.fr/

Abstract. Linked Open Data is an essential part of the Semantic Web. More and more data sets are published in natural languages comprising not only English but other languages as well. It becomes necessary to link the same entities distributed across different RDF data sets. This paper is an initial outline of the research to be conducted on cross-lingual RDF data set interlinking, and it presents several ideas how to approach this problem.

Keywords: Multilingual Mappings, Cross-Lingual Link Discovery, Cross-Lingual RDF Data Set Linkage

1 Motivation

Semantic Web technologies comprise different languages for expressing data as graphs (RDF), describing its organization through ontologies (OWL) and querying it (SPARQL). The Web of Data uses this technology to publish data on the Web. In particular, Resource Description Framework (RDF)[1] - is a standard model for data representation on the Web proposed by W3C[2]. It is designed to represent meta-data about Web resources in the form of triples (Subject, Predicate, Object) and is intended to be processed by machines rather than humans.

The publication of data sets along the Linked Data principles is gaining an increasing importance. The Linked Data Cloud[3] contains data sets from several domains: geographic, media, government, etc. According to the statistics[4], the total number of triples over all 295 data sets reaches 31,634,213,770. Moreover, the Data Hub[5] contains 5034 data sets most of which are publicly available for use. Given this increasing number of the available data sources, one of the key challenges of Linked Data is to be able to discover links across data sets [1]. Interlinking RDF data sets is the process of setting links between related entities. Moreover, given the growth of linked data, automatic methods are necessary to scale. At present, the number of languages[6] of RDF data sets amounts to 474.

[1] http://www.w3.org/TR/rdfprimer/
[2] http://www.w3.org/2001/sw/wiki/Main_Page
[3] http://richard.cyganiak.de/2007/10/lod/
[4] http://wifo5-03.informatik.uni-mannheim.de/lodcloud/state/
[5] http://datahub.io/
[6] http://stats.lod2.eu/languages

Thus, the interlinking problem becomes particularly difficult when entities are described in different natural languages since a simple string comparison of entity labels does not suffice.

Research Question. Our core research problem is to provide automatic reliable methods to link disparate RDF data sets published with labels and literals in various natural languages. Since different URIs can refer to the same real-world object, the focus is on identity links, i.e. a link established between two URIs referring to the same resource. The output of the interlinking process is a set of triples of type: URI owl:sameAs URI.

A reliable method in this context should be understood as a method that links the identical entities across data sets with a high precision and recall. It should also be adaptable to a variety of languages. Though some authors [2] distinguish between multilingual and cross-lingual aspects of matching, we consider them interchangeable.

Research sub-questions to be addressed are as follows:

- Are there monolingual methods adequate for our task? Under what conditions language-dependent methods perform better than language-independent ones?
- What are suitable methods for RDF data set interlinking from Computer Science and Natural Language Processing (NLP) perspectives?
- What method works best for cross-lingual RDF data linking? So far, we plan to identify methods working in broad domains.
- Is there a dependency between families of languages to be linked (Indo-European, Sino-Tibetan, Afro-Asiatic) and the quality of the generated links? This dependency could be traced by changing language pairs of data sets to be interlinked.

The potential contribution of this research is to provide or combine methods to facilitate discovering knowledge across data sets where the same entity is described in different natural languages. Some other cross-lingual applications may benefit from the obtained results: Cross-Language Information Retrieval via Semantic Search engines, Document Classification/Clustering, Question Answering, Machine Translation, to name a few. The cross-lingual mappings obtained as a result of the interlinking process will be shared on the Web for further exploitation by multilingual information access tools in order to facilitate access to knowledge across languages.

In the next section we outline several angles from which the entity linking problem can be looked at.

2 State-of-the-Art

Our research will draw upon the knowledge from different domains: Computer Science, Artificial Intelligence, Natural Language Processing, and Data Mining.

The problem of finding correspondences between entities representing the same world object in distinct data sets has been widely studied in the 1960s in

the context of databases. It is known as instance identification, record linkage or record matching problem. In [3], the authors use the term "duplicate record detection" and provide a thorough survey on the matching techniques. Though the work done in record linkage is similar to our research, it does not contain cross-lingual aspect and RDF semantics.

Our research topic belongs to the area of data linking. String similarity measures [4] and linguistic resources [5] are used to computed the distance between the entities. Another type of approach is to use the features of Linked Data [6].

In the NLP area, the problems of entity resolution, multilingual entity recognition and cross-document co-reference resolution [7] gain a close attention due to their complexity and importance for Information Retrieval, QA, etc. The task is to find out whether the occurrences of a name in different plain natural language texts are the same. There is no general solution to this problem, and the decision whether or not two names refer to the same entity usually relies on contextual clues. One of the differences with the task of finding correspondences between RDF data sets is the limited amount of textual data presented in such data sets which makes it more difficult to calculate similarity measure. Moreover, the RDF graph model and RDF semantics can be of use while elaborating linking strategies.

Recent developments have been made in the field of multilingual ontology matching [8, 10]. Some work has also been done in creating a multilingual ontology known as BabelNet [9]. This resource can be used for word sense disambiguation and is available in RDF format.

To the best our knowledge, the area of multilingual RDF data sets interlinking which could combine both NLP techniques and information from Linked Data has not seen many studies. The current research will attempt to fill this gap.

3 Proposed Approaches

To achieve our goal, we may not invent a new approach but rather combine existing methods and adapt them for RDF data sets in a multilingual context. Below we highlight several commonly known methods to deal with natural language data which may contribute to this goal.

Semi-automatic or automatic linkage heuristics can be appropriate for generating RDF links between heterogeneous data sources. Machine learning techniques can be used to learn how to match entities. The major drawback of supervised learning would be its dependency on availability of training examples (cross-lingual entity links labeled as matching or not), whereas the difficulty for unsupervised learning would be to define a matching threshold.

Given the multilingual nature of the research topic, some applications from NLP are likely to be exploited. For example, Machine Translation can be used to translate one data set into the language of the other set thus attempting to facilitate computation of similarity metrics. Though it is not always true since the results of Machine Translation systems can be far from perfect and introduce

errors decreasing the overall precision. Besides, more than one translation of a particular fact can exist.

Sometimes, a significant part of important information in a text is associated with named entities, for instance, people names, place names, company names. Those might be valuable discriminators when it is necessary to determine whether two documents are about the same entity. Such open-source free text analysis toolkits as GATE[7] and OpenNLP[8]can be used for Named Entity Recognition and Information Extraction tasks.

4 Planned Research Methodology

The main aim of this work is to find reliable and scalable methods and develop tools for linking different URIs used to identify the same resource represented in multiple natural languages and located in different RDF data sets.
To achieve this aim, the research will go through the following steps:

- Synthesize the work done in the research field
- Select the acceptable RDF data sets
- Deal with a problem of partially built data sets
- Explore the semi-automatic and automatic techniques for RDF interlinking
- Decide what methods to choose and how to combine them
- Run experiments on actual data sets
- Evaluate and analyze the obtained empirical results

The research procedure can be summed up as follows:

- Internet-based data collection method will be used to obtain RDF data sets.
- Once the research methods are refined, the experiments will be conducted in order to obtain RDF links between corresponding entities.
- Since we will attempt to automate the linking process as much as possible, standard statistical measures will serve for evaluation. As a starting point, the results of the best multilingual ontology matcher [10] with F-measure = 18% could be considered as a baseline. As of today, there is no official benchmark for doing evaluation. This poses difficulties to objective evaluation of method effectiveness. The problem could be addressed in several ways. One way would be to create reference links manually. The other way would be to exploit existing links between knowledge bases (for example, multilingual DBpedia): first, the existing links are deleted, then the methods are applied and the obtained links are compared against the initially deleted links. Another possible direction is to elaborate a task-oriented evaluation with a well-defined application for evaluating the correctness of the obtained links.

[7] http://gate.ac.uk/
[8] http://opennlp.apache.org/

5 Schedule

Time allowance to complete the proposed research is 36 months. Below we present a rough schedule with important milestones for every 6 months.
M0-M6: Attend courses and research seminars; bibliographic study
M0-M12: Finalize research methodology, collect corpora and configure software for experiments
M6-M18: Propose problem solutions and conduct preliminary experiments
M18-M24: Analyze results and prepare publication
M24-M30: Generalize results and conduct further experiments
M26-M32: Write up and prepare publications
M32-M36: Send for review and correct final version of thesis

References

1. Ferrara, A., Nikolov, A. and Scharffe, F.: Data Linking for the Semantic Web. Int. J. Semantic Web Inf. Syst. 7(3), 46-76 (2011)
2. Spohr, D., Hollink, L., Cimiano, P.: A machine learning approach to multilingual and cross-lingual ontology matching. In: 10th ISWC, pp. 665-680(2011)
3. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering. 19(1), 1-16 (2007)
4. Winkler, W.: Overview of record linkage and current research directions. Tech. Rep. No.2006-2. Statistical Research Division. U.S. Census Bureau.
5. Scharffe, F., Liu, Y., Zhou, C.: RDF-AI: an architecture for RDF datasets matching, fusion and interlink. In: Workshop on Identity and Reference in Knowledge Representation, IJCAI, Pasadena, CA, USA (2009)
6. Hu, W., Chen, J., Qu, Y.: A self-training approach for re-solving object coreference on the semantic web. In: Proc. 20th International World Wide Web Conference (WWW 2011), pp. 87-96, Hyderabad, India (2011)
7. Bagga, A., Baldwin, B.: Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In: Proc. 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, pp. 79-85, (1998)
8. Meilicke, C., Castro, R.G., Freitas, F., van Hage, W.R., Montiel-Ponsoda, E., de Azevedo, R.R., Stuckenschmidt, H., Svb-Zamazal O., Svtek, V., Tamilin, A., Trojahn, C., Wang, S.: MultiFarm: A Benchmark for Multilingual Ontology Matching. Journal of Web Semantics. 15, 62-68 (2012)
9. Navigli, R., Ponzetto, S.: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence. 193, 217-250 (2012)
10. Meilicke, C., Trojahn, C., Sváb-Zamazal, O., Ritze, D.: Multilingual Ontology Matching Evaluation - a First Report on Using MultiFarm. In: Proc. 2d International Workshop on Evaluation of Semantic Technologies, pp.1-12, Heraklion, Greece (2012)