HAL open science

Interlinking RDF data in different languages

Tatiana Lesnikova

To cite this version:

Tatiana Lesnikova. Interlinking RDF data in different languages. 4th workshop on Terminology & Ontology: Theories and applications (TOTh), Dec 2014, Bruxelles, Belgium. hal-01180921

HAL Id: hal-01180921

https://hal.science/hal-01180921v1

Submitted on 28 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Interlinking RDF Data in Different Languages

Tatiana Lesnikova

University of Grenoble Alpes & Inria, Grenoble, France
tatiana.lesnikova@inria.fr
http://exmo.inria.fr/

Keywords: semantic web, cross-lingual resource discovery, owl:sameAs, multilingual instance matching

The Semantic Web provides technologies such as the Resource Description Framework (RDF) for representing data on the web. The number of data sets published in RDF is growing rapidly. To connect a variety of data across RDF data sets, they should be interlinked. However, resources can be described in different natural languages. Such publishers as the French National Library [1], the Spanish National Library [2], the National British Museum[1] make their data available using RDF model in their own language. There also exist encyclopedias in RDF: DBpedia in multilingual versions is a structured information from Wikipedia; XLore database [3] is an effort to publish the Chinese encyclopedias (Baidu Baike and Hudong Baike) in RDF. The Europeana Project[2] aims at bringing together descriptions of cultural artifacts from European cultural institutions. This is done by harvesting the metadata of its data providers. The descriptions of these artifacts can be in different languages. The importance of tackling multilingualism in the semantic web has been highlighted in [4].

Problem description. Given two RDF data sets with resources described in different languages, the same entity represented in different data sets has to be identified. At the instance level, the values of properties are in different languages, which makes it harder to merge data about the same entity from different sources.

The goal of our research is to identify the same entities across multilingual RDF data sets and link them by owl:sameAs links. For this purpose, we are developing an approach which represents RDF entities as text documents and then compare them. We apply standard Natural Language Processing (NLP) techniques (document preprocessing, term weights, similarity measures) on our data. We particularly explore two strategies presented below.

1 Applying Machine Translation (MT) in cross-lingual RDF data interlinking

In our first experiment on interlinking RDF resources described in English and Chinese languages [5], we used a machine translation approach (see Figure 1).

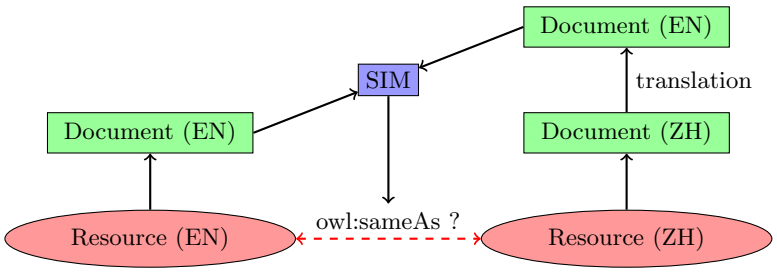

Fig. 1: Linking Process. Interlinking of DBpedia in English and Xlore in Chinese languages.

Our objective was to test parameters (levels of information per resource, various document preprocessing, terms weights as Term Frequency and TF*IDF, three ways of extracting links)

[1] http://collection.britishmuseum.org/
[2] http://europeana.eu/portal/

and find out the most efficient ones. We represented resources as text documents and translated the Chinese data into English using a statistical MT system. The translation was done only in one direction. Documents were represented as vectors using two weighting schemes, then cosine similarity was computed. Similarity between documents was taken for similarity between resources. As a result, we determined that the method can identify most of the correct matches. Using minimum information in a resource description combined with TF*IDF, we obtained F-measure over 95%. We also showed that the mismatches were likely to occur between entities belonging to the same category, which means that our method can work without prior ontology matching.

2 Using references to external multilingual resources

In our second experiment, we focused on a multilingual resource-based approach for instance matching. The availability of such open information sources as Wikipedia with more that 280 language editions[3], DBpedia – a nucleus of the Linked Open Data with multiple language versions, BabelNet [6] – a semantic network which connects concepts and named entities in 50 languages – can be a valuable aid for our task.

We took BabelNet as an external multilingual knowledge base and have conducted an experiment on the Chinese-English language pair. We created several RDF data sets with different characteristics (presence or absence of rdfs:label property and/or non-matching entities), represented text documents as vectors of BabelNet identifiers and applied the same similarity measures. Overall, our preliminary results were lower with F-measure of 62% for the best result. We are working on improving them.

Conclusion

Multilingual resources (machine translation systems, dictionaries, knowledge-bases, encyclopedias) play an important role in a cross-lingual data interlinking task and are valuable tools for multilingual information processing. Linking entities in a multilingual context relies heavily on such resources. The multilingual resource interlinking can help to uncover the potential of vast amounts of linked open data and facilitate knowledge discovery across language barriers.

Acknowledgments. Research is partially supported by the Lindicle[4] (12-IS02-0002) project in cooperation with the Tsinghua University, China.

References

1. Simon, A., Wenz, R., Michel, V., and Mascio, Adrien Di: Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF (French National Library). In: ESWC, volume 7882 of Lecture Notes in Computer Science, pp.563–577 (2013)
2. Vila-Suero, D., and Villazón-Terrazas, B., and Gómez-Pérez, A.: datos.bne.es: A library linked dataset. Journal of Semantic Web. 4(3), 307–313 (2013)
3. Wang, Z., Li, J., Wang, Z., Li, S., Li, M., Zhang, D., Shi, Y., Liu, Y., Zhang, P., Tang, J.: XLore: A Large-scale English-Chinese Bilingual Knowledge Graph. In: International Semantic Web Conference (Posters & Demos), Vol. 1035 of CEUR Workshop Proceedings, pp. 121–124. CEUR-WS.org, (2013)
4. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gômez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the Multilingual Web of Data. Journal of Web Semantics. 11, 63–71 (2012)
5. Lesnikova, T., David, J., Euzenat, J.: Interlinking English and Chinese RDF Data Sets Using Machine Translation. In: Proc. 3rd ESWC workshop on Knowledge discovery and data mining meets linked open data (Know@LOD), Vol. 1243 of CEUR-WS.org, (2014)
6. Navigli, R., Ponzetto, S.: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence. 193, 217-250 (2012)

[3] http://en.wikipedia.org/wiki/List_of_Wikipedias
[4] http://lindicle.inrialpes.fr/