NLP for Interlinking Multilingual LOD

Tatiana Lesnikova

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

Abstract. Nowadays, there are many natural languages on the Web, and we can expect that they will stay there even with the development of the Semantic Web. Though the RDF model enables structuring information in a unified way, the resources can be described using different natural languages. To find information about the same resource across different languages, we need to link identical resources together. In this paper we present an instance-based approach for resource interlinking. We also show how a problem of graph matching can be converted into a document matching for discovering cross-lingual mappings across RDF data sets.

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

1 Problem Statement

Due to the Resource Description Framework (RDF), the information on the Web can be turned from the unstructured mass into the structured data represented in the form of triples. The Linked Open Data (LOD) cloud containing billions of triples is constantly growing. Since data sets are created independently, there can be several Uniform Resource Identifiers (URIs) denoting the same entity across different RDF data sets. As a result, one needs to address the problem of entity resolution: identify and interlink the same entity across multiple data sources.

The RDF syntax is relatively simple and unambiguous: RDF = graph + identifiers (labels). This is what the identification of resources can be based on. However, this problem can become particularly difficult when there are multilingual elements in a graph as a simple string matching technique is doomed to fail. Hence, specific Natural Language Processing (NLP) techniques must be considered.

Our research problem is to find out methods for linking the same resource located in several RDF data sets and described in various natural languages and study the impact of available NLP techniques on the interlinking procedure.

2 Relevancy

Internet is a multilingual system, and we believe that it will continue to accommodate a diversity of natural languages despite the development of the Semantic Web. Even though there are many resources in English, some other languages occupy a decent portion of the Web space as well (see Internet world users by language statistics [1]). And we expect the necessity to tackle the multilinguality problem to persist. There are many resources that could be interlinked. At present, the number of languages [2] of RDF data sets amounts to 503.

The importance of cross-lingual mappings has been discussed in several works [1–3].

Recently a Best Practices for Multilingual Linked Open Data Community Group [3] has been created to elaborate a large spectrum of practices with regard to multilingual LOD.

Availability of the cross-lingual links is imperative for several neighboring research areas. For example, to overcome the problem of ontology heterogeneity, some research has been done on monolingual ontology integration based on instances interlinked by owl:sameAs [4]. If owl:sameAs links could be provided between instances expressed in different languages, other experiments on integrating underlying ontologies could be conducted.

The owl:sameAs links between instances can be also valuable in other applications such as Question Answering over multilingual structured knowledge-base [5] since a system can take advantage of the information presented in a language different from a language that is being queried.

Thus, the growing number of data sources in RDF format with multilingual labels and the importance of cross-lingual links for other Semantic Web applications motivate our interest in cross-lingual link discovery.

3 Related Work

The problem of searching for the same entity across multiple sources has been investigated in several research fields. In database community, it is known as instance identification, record linkage or record matching problem. In [6], 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.

In the field of Information Retrieval (IR), within the framework of the Cross-Language Evaluation Forum (CLEF)[4], the Web People Search Evaluation Campaigns (2007-2010)[5] focused on the Web People Search and person name ambiguity on Web pages and aimed at building a system which could estimate the

[1] http://www.internetworldstats.com/stats7.htm
[2] http://stats.lod2.eu/languages
[3] http://www.w3.org/community/bpmlod/
[4] http://www.clef-initiative.eu/
[5] http://nlp.uned.es/weps/weps-3

number of referents and cluster Web pages that refer to the same individual into one group. The research was performed on monolingual data.

Cross-lingual entity linking has been addressed in Knowledge Base Population track (KBP2011)[7] within the Text Analysis conference. The task is to link entity mentions in a text to a knowledge base (Wikipedia). If entity mentions are not in KB, they should be clustered into a separate group. Experiments were done both on monolingual (English) and cross-lingual data (Chinese to English). Authors in [8] used both language-independent and translation-based methods.

In contrast to the research outlined above, we aim at providing insights into the problem of cross-lingual interlinking from the point where data are already in RDF format, and we can vary different parameters in order to determine their impact on the interlinking operation.

In the Semantic Web, interlinking resources that represent the same real-world object and that are scattered across multiple Linked Data sets is a widely researched topic. Within the Data Interlinking track (IM@OAEI 2011), several interlinking systems have been proposed [9–13]. All of the systems were evaluated on monolingual data sets. Recent developments have been made also in multilingual ontology matching [14, 15].

To the best of our knowledge, there is no interlinking system specifically designed to link RDF data sets with multilingual labels.

4 Research Questions

The goal of our work is to provide methods to link interrelated resources across multilingual RDF data sets. For now, we restrict ourselves to owl:sameAs link [16] as it is a classical type of link that is usually established, and it is also important for tracking information about the same resource across different data sources. Given two RDF data sets with URIs and literals in different natural languages, the output will be a set of triples of type URI1 owl:sameAs URI2.

Our general *research question* is: To what extent is it possible to interlink data sets in different languages? To answer this question, within the framework that we describe in the Proposed Approach section, we need to explore which parameters influence this task. More specifically:

1. How to represent entities from RDF graphs?
   - What is the optimal distance for collecting language elements in traversal?
   - Is it necessary to preserve the structure of the graph in a virtual document by weighting the path length?
2. How to make entities described in different natural languages comparable?
   - What are the most appropriate Machine Translation techniques (rule-based, statistical, hybrid)?

- What is the impact of translating one language into another or pivot language?
- How does the output of similarity measures vary according to the context?

All these parameters will be studied with respect to specific contexts (language pairs, data set types, amount of textual data available). We also plan to experiment with graph matching techniques to see the difference with a translation-based approach. Apart from Machine Translation, we will explore techniques used for word alignment, thesaurus-based word sense disambiguation, multilingual document ranking, and mapping to multilingual lexicons.

5 Hypotheses

We introduce several hypotheses that we would like to test in our research.

1. If two URIs denote the same real-world object, the descriptions of the properties of this object should overlap with each other.
2. If descriptions are in different natural languages, then NLP techniques could help to decrease uncertainty across a set of resources.
3. If the descriptions of an entity overlap significantly, the similarity between them will be higher than between other entities.
4. If the degree of similarity depends on the available language context for each entity, then the more language data there are, the better will be the matching results.
5. If language data can be taken from two sources in RDF graph: property names and literals; then literals are more important since they are more informative.

6 Proposed Approach

Due to the presence of natural language terms in RDF graphs, we adopt a language-oriented approach.

The proposed approach includes several steps (see Figure 1).

1. Given two data sets with a resource representation in different natural language, extract language data for each URI. Thus, we create a "virtual" document for each URI.
2. Compare virtual documents in pairs from both sets.
3. Find the maximum similarity between two representations of the resource.
4. Establish an owl:sameAs link between the two most similar representations.

One should mind the following aspects of this approach:

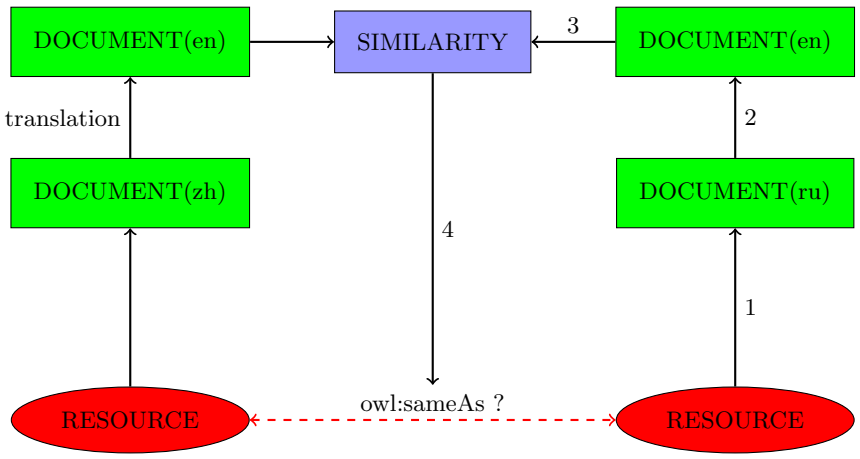

Fig. 1. Linking Process. Resources are described in Chinese and Russian languages and then translated into English.

– The idea of creating a "virtual" document has been employed in ontology matching [17]. The intuition of converting a graph into a document representation is that even though the taxonomy (structure) of graphs can be similar, the possibility to distinguish between two different things and identify the identical ones relies on their comparison. Thus, it is important to take into account lexical elements in a graph.

– Once we have documents representing resources, we need to decide how to define similarity between these resources. Similarity between documents can be taken for similarity between resources. Since we have documents in different languages, we can experiment with different types of Machine Translation (statistics-based, rule-based, hybrid). To estimate which strategy yields a better result, we will run our system by changing the translation component iteratively. Significant difference in results may signal which translation type is more beneficial. To enhance scalability, it would be interesting to translate the whole source corpus once and not to translate each label again and again. This would also allow for more contextual translation. The choice of translation techniques can also depend on the language combinations, for example, for rare languages, for which there does not exist enough parallel corpora, dictionary-based approaches might help.

– At the resource comparison step, it is important to reduce a number of possible comparisons for the sake of time-efficiency. For example, only comparisons between certain entity types are allowed. In case of using Supervised Machine Learning, the problem of training data is the most prominent one since there has been no official benchmark. And creating a generic training set for a heterogeneous amount of Linked Data seems very unrealistic. Then,

instead of training, it would be interesting to test clustering algorithms and find appropriate parameters for identity resolution.

- There are many techniques to compute similarity. A broad overview of them is given in [18]. We will use a vector space model [19] to represent terms in a "virtual" document as vectors of features. The choice of particular similarity measures is yet to investigate. When terms are in different languages, document similarity fails. Some similarity measures perform better on long texts. After transformation of "virtual" document into vectors, similarity metrics (e.g. Cosine, Euclidean) can be computed.

- A virtual document per URI shall contain language data in proximity to a given URI. The hypothesis is that the more textual data we have to characterize a resource, the easier it would be to identify the identical ones.

There are some complications as to textual data. Two scenarios are possible:

- URI can be looked up and the textual data extracted (as in case of DBpedia)
- No extra textual data are available per URI except the data in a graph itself.

To overcome this lack of context for a particular resource, we propose to browse a graph up to n+1 hops from the URI under investigation and collect data along the way. The data carriers are property names and literals. Thus, a virtual document for a particular URI will be the accumulation of data gathered during graph traversal.

This way of collecting a "profile" for a resource entails a question: does the difference of two graph structures affect the results of interlinking? On the one hand, taking into account the success of statistical machine translation based on statistical modelling and probabilities, the order of words is not always that important. It would be interesting to see whether it holds for RDF interlinking as well. On the other hand, we can try to preserve the order of collected properties and literals in a virtual document by putting weight for each language element. The further it is from the URI at question, the lower the weight. Term weight can be assigned by computing $termfrequency$ in a document or distribution of terms across a collection of documents known as $inverse\ documentfrequency$ (IDF). Terms that appear in few documents can be discriminative with regard to the rest of the documents. Combination of both TF x IDF is widely used in vector space models.

Once virtual documents are collected from both graphs, the documents will be compared and results evaluated.

7 Reflections

We believe that we can succeed in finding the solution for our research topic because we plan to put our research on a solid foundation and combine different methods to achieve the task. In traditional Web, there has been much

work done on multilingual NLP, i.e. language identification, machine translation, cross-language information retrieval. We are going to conduct series of experiments and see what works and how we can improve what does not work. This would allow us to preserve only the best practices and finally crystallize a solution to the problem. The author of this research proposal is also guided by the specialists in the domain that will contribute to the right choice of the research direction.

8 Evaluation

Evaluation means comparing the retrieved links against some reference. Standard measures usually serve for evaluation of an interlinking system (Precision, Recall, F-measure). The biggest challenge for evaluating our system is the absence of standard benchmark tests. As described in [20], there are several ways to go about this challenge. One of them would be to rely on the existing links between resources in DBpedia. This could be considered as a good alternative if not yet another hurdle: the existing interlanguage links can be inaccurate [21]. So, in our research we plan to experiment with different evaluation settings: we may experiment only with bi-directional links and/or study transitivity in order to ensure the correctness of test cases. The English, French, Russian versions of DBpedia [6] and Baidu Baike in Chinese [22] will be used for our experiments. We will also try to identify types of entities to focus on.

References

1. Gracia, J., Montiel-Ponsoda, E., Gômez-Pérez, A.: Cross-lingual Linking on the Multilingual Web of Data. In: Proc. of the 3rd Workshop on the Multilingual Semantic Web (MSW 2012) at ISWC 2012, Boston (USA), CEUR-WS ISSN 1613-0073, vol. 936 (2012)
2. 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)
3. Buitelaar, P., Choi, K.-S., Cimiano, P., Hovy, H. E.: The Multilingual Semantic Web (Dagstuhl Seminar 12362). Dagstuhl Reports 2(9), pp.15–94 (2012)
4. Zhao, L., Ichise, R.: Instance-Based Ontological Knowledge Acquisition. In: Proc.10th International Conference, ESWC 2013, Vol. 7882, pp.155–169. LNCS, Springer Berlin Heidelberg (2013)
5. Cabrio, E., Cojan, J., Gandon, F., and Hallili, A.: Querying multilingual DBpedia with QAKiS. In: Proc. 10th International Conference, ESWC 2013. Demo paper. Montpellier, France (2013)
6. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering. 19(1), 1–16 (2007)
7. Ji, H.,Grishman, R., Dang, H.T.: An Overview of the TAC2011 Knowledge BAse Population Track. In: Proc. Text Analytics Conference (TAC2011) (2011)

[6] http://dbpedia.org/About

8. Monahan, S., Lehmann, J., Nyberg, T., Plymale, J., Jung, A.: Cross-Lingual Cross-Document Coreference with Entity Linking. In: Proc. TAC2011 (2011)
9. Ngonga Ngomo, A.-C., Auer, S.: LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. IJCAI, pp. 2312–2317 (2011)
10. Nguyen, K., Ichise, R., Le, B.: SLINT: A Schema-Independent Linked Data Interlinking System. In: Proc. of the 7th International Workshop on Ontology Matching, pp.1–12 (2012)
11. Volz, J., Bizez, C., Gaedke, M., and Kobilarov, G.: Discovering and maintaining links on the web of data. In: Proc. of ISWC' 09, Springer-Verlag Berlin, Heidelberg, pp. 650–665, 2009.
12. Araújo, S., Hidders, J., Schwabe, D., Arjen, P. de Vries: SERIMI - Resource Description Similarity, RDF Instance Matching and Interlinking. CoRR, abs/1107.1104 (2011)
13. Niu, X., Rong, S., Zhang, Y., and Wang, H.: Zhishi.links results for OAEI 2011. In: Proc. of ISWC' 11 6th Workshop on Ontology Matching, pp. 220–227 (2011)
14. Meilicke, C., Trojahn, C., Sváb-Zamazal, O., Ritze, D.: Multilingual Ontology Matching Evaluation - a First Report on Using MultiFarm. In: Proc. of the 2d International Workshop on Evaluation of Semantic Technologies, pp.1–12, Heraklion, Greece (2012)
15. Meilicke, C., Castro, R.G., Freitas, F., van Hage, W.R., Montiel-Ponsoda, E., de Azevedo, R.R., Stuckenschmidt, H., Sváb-Zamazal O., Svátek, V., Tamilin, A., Trojahn, C., Wang, S.: MultiFarm: A Benchmark for Multilingual Ontology Matching. Journal of Web Semantics. 15, 62–68 (2012)
16. Halpin, H., Hayes, J. P.: When owl: sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In: Proc. of the Linked Data on the Web Workshop (LDOW2010), Raleigh, North Carolina, USA, April 27, 2010, CEUR Workshop Proceedings, ISSN 1613-0073, online http://ceur-ws.org/Vol-628/ldow2010_paper09.pdf
17. Qu, Y., Hu, W., Cheng, G.: Constructing virtual documents for ontology matching. In: Proc. of the 15th International Conference of World Wide Web, pp.23–31 (2006)
18. Euzenat, J. and Shvaiko, P.: Ontology Matching. Springer-Verlag, Heidelberg (2007)
19. Salton, G.: The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs, NJ (1971)
20. Lesnikova, T.: Interlinking Cross-Lingual RDF Data Sets. In: Proc. 10th International Conference, ESWC 2013, Vol. 7882, pp. 671-675. LNCS, Springer Berlin Heidelberg (2013)
21. Rinser, D., Lange, D., Naumann, F.: Cross-lingual entity matching and infobox alignment in Wikipedia. Information Systems, 38 (6), pp. 887-907 (2013)
22. Wang, Z., Wang, Z., Li J., Pan, J. Z.: Knowledge extraction from Chinese wiki encyclopedias. Journal of Zhejiang University - Science C 13(4): 268-280 (2012)