Interlinking English and Chinese RDF Data Using BabelNet

Tatiana Lesnikova
Univ. Grenoble Alpes & INRIA
Grenoble, France
tatiana.lesnikova@inria.fr

Jérôme David
Univ. Grenoble Alpes & INRIA
Grenoble, France
jerome.david@inria.fr

Jérôme Euzenat
INRIA & Univ. Grenoble Alpes
Grenoble, France
jerome.euzenat@inria.fr

ABSTRACT

Linked data technologies make it possible to publish and link structured data on the Web. Although RDF is not about text, many RDF data providers publish their data in their own language. Cross-lingual interlinking aims at discovering links between identical resources across knowledge bases in different languages. In this paper, we present a method for interlinking RDF resources described in English and Chinese using the BabelNet multilingual lexicon. Resources are represented as vectors of identifiers and then similarity between these resources is computed. The method achieves an F-measure of 88%. The results are also compared to a translation-based method.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing, Dictionaries; I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic networks; E.2 [**Data Storage Representations**]: Linked representations

General Terms

Semantic Web, Cross-lingual Data Interlinking

Keywords

Cross-lingual Instance Linking, Cross-lingual Link Discovery, owl:sameAs

1. INTRODUCTION

Linked Data enables the extension of the Web based on Semantic Web technologies. RDF (Resource Description Framework) is a W3C data model according to which a resource is described by triples (subject, predicate, object). The RDF statements form a directed labeled graph where the graph nodes represent resources and the edges represent relations between these resources. A set of statements about a resource constitutes a description set which contains certain characteristics of a resource and thus can ground the resource "identity".

Knowledge can be expressed in different languages. DBpedia[1] provides a semantic representation of Wikipedia in which multiple language labels are attached to the individual concepts. It has become the nucleus for the Web of Data. Though there are interlingual links between different language versions of Wikipedia, there are knowledge bases in other languages which are not interlinked. For example, XLore [8] is an RDF Chinese knowledge base which provides a semantic representation of national knowledge sources (Baidu baike, Hudong baike).

Cross-lingual interlinking consists in discovering links between entities across knowledge bases of different languages. It is particularly difficult due to several reasons: (1) the structure of graphs can be different and the structure-based techniques will not be much of help; (2) even if the structures are similar to one another, the properties themselves and their values are expressed in different natural languages. In this regard, we adopt a Natural Language Processing (NLP) approach to address the problem of finding the same object described in two different languages. Our hypothesis is that if two resources denote the same real-world object, then the descriptions of these resources should overlap with each other.

In this paper, we propose an instance interlinking method based on a multilingual lexicon which serves as a pivot language in order to make two resources comparable. We describe an experiment on interlinking resources with English and Chinese labels across data sets and compare it with a translation-based method. Given two RDF data sets, our goal is to find the identical resources and to interlink them with owl:sameAs link. This type of link is important for tracking information about the same resource across different data sources. The paper answers the following questions:

- Is a multilingual lexicon an appropriate medium to identify resource in two different languages?
- What method performs better: a method based on translation technology or multilingual lexicon?

The remainder of the paper is structured as follows. Section 2 presents related work on interlinking methods. Section 3 describes the proposed approach based on multilingual lexicon. Section 4 describes a corpus used in the experiments and evaluation scenarios. Results of the experiments are shown in Section 5. We outline our contributions and propose directions for future work in Section 6.

[1]http://wiki.dbpedia.org

DOI: http://dx.doi.org/10.1145/2682571.2797089
To appear in Proc. ACM DocEng, pp39–42, Lausanne, Switzerland, 2015.
Author version..

2. RELATED WORK

The problem of finding the same object across heterogeneous data sources has many names: duplicate matching (deduplication), record linkage (in the database field), entity matching, entity resolution, object identification, instance matching. In the Semantic Web, data interlinking is the task of finding the same entity within different RDF graphs. The challenges for multilingual Web of data and linking procedures have been highlighted in [2]. In a cross-lingual context, interoperability involves linking identical resources described in different languages. A comprehensive survey of techniques for data linking can be found in [6]. The use of string matching is a widespread technique to identify similarity between entities, however, in a cross-lingual context, string matching algorithms will not work.

Datasets can be described by ontologies. Even if the ontologies are in the same language, the difference in granularity of categories can complexify the process of ontology matching. Recent developments have been made also in cross-lingual ontology matching [4]. A common approach to break the natural language barrier consists in transforming a cross-lingual problem into a monolingual one by translating the elements of one ontology into the language of the other ontology using machine translation [1]. After translation, monolingual matching strategies are applied.

In previous experiments [3], we described an interlinking method which has been relying both on language elements in a graph and machine translation. In this method, given two RDF graphs with resources described in different languages, each resource has been represented as a virtual document containing textual information from n neighboring nodes. Once constructed, these documents have been translated and standard text processing techniques have been applied. The similarity computed between documents is taken for similarity between resources. The pair of documents with the highest similarity score has been considered as a correspondence between identical resources.

Furthermore, instead of translation, it is possible to use multilingual lexical resources to compute semantic relatedness between entities. In the next section, we detail how an external multilingual resource can be used in interlinking RDF data across languages.

3. CROSS-LINGUAL INTERLINKING METHOD

We assume that resources in RDF are described with labels in different natural languages: properties and their values are usually natural language words. We adopt a linguistic interlinking approach where the textual description of a resource is very important: the similarity score highly depends on the overlapping text.

The framework that we designed for interlinking cross-lingual RDF resources is depicted in Figure 1, extending the process presented in [3].

In the present approach, we use a multilingual lexicon which serves as a basis for resource comparison. The interlinking method is schematized in Figure 2.

In particular, the method is the following:

1. Constructing a **Virtual Document** per resource. Due to the graph structure of RDF, we collect literals up to a specific distance (that we call level). The triples of an RDF graph can have simple strings (literals) as an object which serve as a descriptor for a subject. We collect literals from all resource properties, the names of the properties themselves are not considered. In the example of Figure 3, the subject is "dbpedia:Lucerne" which has several literals, e.g., the label "Lucerne". These collected literals will constitute the body of a virtual document. We work with level 1 and 2 only.

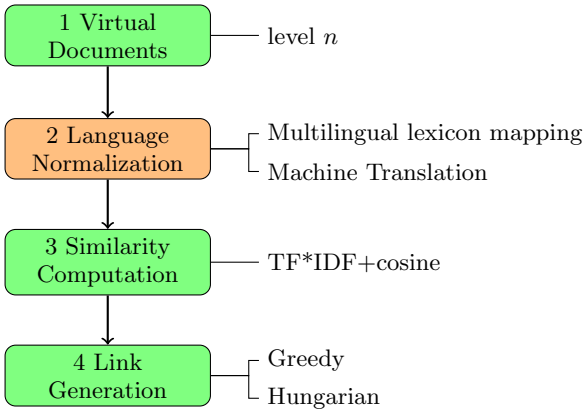

Figure 1: Framework for Cross-lingual RDF Interlinking.

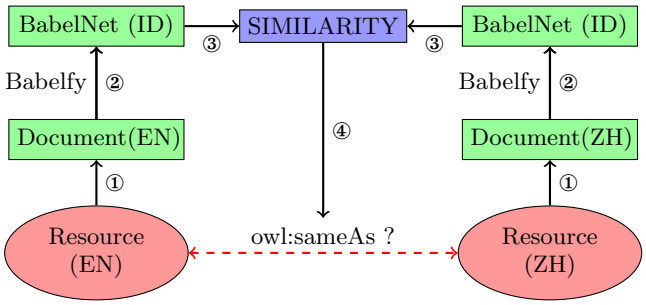

Figure 2: Interlinking Method Using Multilingual Lexicon. Multilingual terms are mapped to a common identifier. Similarity is computed between identifiers. Numbers correspond to the steps of the method.

2. Replacing document terms by identifiers from a **Multilingual Lexicon** in order to project the words of each language onto the same semantic space. At this step, we represent original documents as vectors of identifiers (IDs). A corresponding identifier (ID) is retrieved for a term. An identifier stands for a sense of a term and very often there are many senses (IDs) per term. If more that one sense exists, word sense disambiguation techniques shall be applied in order to select the best sense. The terms not found in a multilingual lexicon are discarded and we do not work with them in our experiments. To compute semantic relatedness, multilingual lexical knowledge resources can be used, e.g., BabelNet [5] or DBnary [7].

3. **Computing Similarity** between documents. We use a standard term weighting scheme (TF*IDF) and apply cosine similarity. These are classical techniques for finding similar documents, moreover, they showed good performance in our previous experiments. The

output of this step is a set of similarity values between pairs of virtual documents.

4. **Generating Links** between identical resources. At this stage, an algorithm extracts links on the basis of the similarity between documents. We use the Hungarian or greedy methods to extract links.

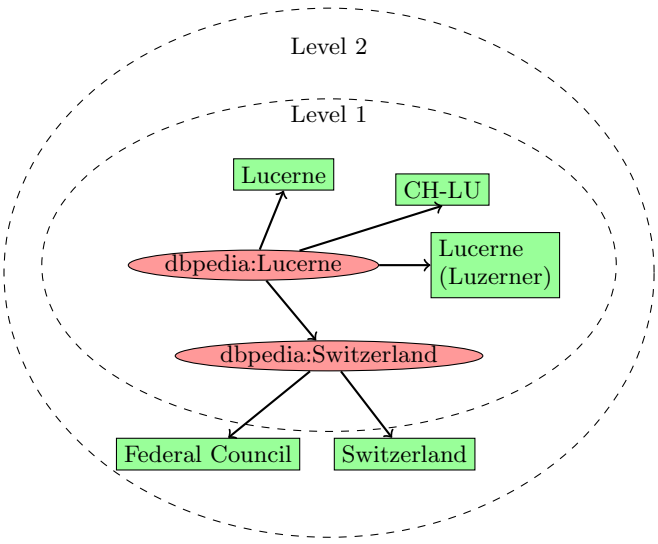

Figure 3: Creation of Virtual Documents by Levels.

4. EVALUATION SETUP

Our goal is to evaluate how the method, described in the previous section, works and what parameters are important. We particularly focus on four parameters: the presence or absence of non-matching entities in a data set, the presence or absence of rdfs:label property values in a virtual document, the amount of text in a virtual document per resource and the link extraction mechanism. We also evaluate the suitability of multilingual lexicon for identifying identical resources.

4.1 RDF Data

The experiment has been conducted on two separate RDF data sets with resources represented in English and Chinese respectively. Thus, the data consist of the English and Chinese parts. For the English part, we used DBpedia 3.9[2], for the Chinese part – Xlore.org[3]. We restricted our experiment to named entities, e.g, presidents, sportsmen, geographical places.

The original data set is the same as described in [3], however we have enhanced it in several aspects. The Chinese data has already been linked to the English version of DBpedia and we used a list of owl:sameAs links as our reference link set at the evaluation step. Two datasets have been constructed:

- Original set: contains 100 entities in one-to-one correspondence in English and Chinese languages.
- Original set + noise: we added 10 entities into each language side which do not have a match in the other language. This has been done in order to observe how similarity works when entities do not have matches.

The Chinese Xlore data set has been already linked to the English DBpedia. We used reference links (existing owl:sameAs links between resources) in order to select a list of entities per category. We selected entities that appeared in a reference link set and contained textual information at both levels and in both languages. The result of this selection is a relatively clean corpus which contains textual description of resources at both levels. This allowed us to test the level at which the performance is better. Entities used as noise are entities which have been present only in one language side and have been selected from the same categories as entities from the original set.

Each of these datasets contains virtual documents of two kinds: with an rdfs:label property value or without it. Thus, we have two variations of each dataset per language: Label and NoLabel.

Since we are linking named entities, an rdfs:label property value is usually a name of the entity which can be highly discriminative. By constructing a virtual document without this property value, we estimate the importance of this element in a resource description. The average number of words in virtual documents of the Original set is 230 at level 1 and 2100 at level 2 for the English language, the numbers do not vary much when noise is added. No such statistics is available for Chinese since we do not use Chinese tokenization (it is done at lexicon-mapping step by Babelfy).

4.2 Experimental parameters

Multilingual lexicon mapping.

We use BabelNet 2.5.1 which is a multilingual lexicon which connects concepts and named entities in a large network of semantic relations called synsets. Each synset represents a given meaning and contains synonyms which express that meaning in a range of different languages. Since many terms can have several synsets, we also made use of Babelfy 0.9 [4] in order to retrieve the best meaning per term. By design, Babelfy had a limit of 3500 characters for input text, so we had to cut documents at level 2 only. The impact of this is that we missed additional textual information which could have been useful for similarity computation.

Machine translation.

We also apply machine translation on the experimental data. We translate virtual documents using Machine Translation in order to transform documents into the same language. We use Bing Translator[5] to translate Chinese documents into English. Once the documents are translated, we preprocess data to prepare it for similarity computation. Virtual documents are treated as "bags of words", and we use standard NLP preprocessing techniques: transform cases into lower case + tokenize + filter stop words. Once the documents are preprocessed, we apply TF*IDF and cosine similarity.

[2]http://wiki.dbpedia.org/Downloads39
[3]http://xlore.org/index.action
[4]http://babelfy.org/
[5]https://www.bing.com/translator/

Table 1: Comparison of MT and BabelNet Methods. Similarity between Entities Using TFIDF. The numbers represent precision (P), recall (R) and F-measure (F) for the Hungarian extraction method.

	Hungarian	Machine Translation						BabelNet					
		level 1			level 2			level 1			level 2		
		P	F	R	P	F	R	P	F	R	P	F	R
Label	**Original set**	1	**1**	1	0.94	0.94	0.94	0.88	**0.88**	0.88	0.83	0.83	0.83
	Original set + noise	0.9	0.94	0.99	0.83	0.87	0.91	0.73	0.76	0.80	0.7	0.73	0.77
NoLabel	**Original set**	0.93	0.93	0.93	0.92	0.92	0.92	0.81	0.81	0.81	0.78	0.78	0.78
	Original set + noise	0.8	0.84	0.88	0.78	0.82	0.86	0.71	0.74	0.78	0.65	0.68	0.71

5. RESULTS

In the current evaluation, we have compared the results obtained using both methods: MT-based and BabelNet, see Table 1. We have compared the results using two popular assignment algorithms: the Hungarian and greedy. The best results have been achieved by the Hungarian algorithm so we do not report the results of the greedy one. The best results are obtained at level 1 on data sets with the rdfs:label property. Results at level 2 decrease for both algorithms: this is because information at level 2 becomes less discriminative and more noisy. Results are also lower when non-matching entities are added. In general, the translation approach outperformed the approach based on multilingual lexicon. This might be due to the better development of MT capability and unavailability of identifiers for some terms as well as errors in disambiguation in BabelNet. Since the terms not found in BabelNet have been discarded (as per step 2 Section 3), we know neither the nature of the missing terms nor the distribution of the number of missing terms per entity. If missing terms are preserved, the absence of identifiers may be compensated by translating those terms using machine translation. The results at level 2 may have been affected by the input text limit of Babelfy.

6. CONCLUSIONS

With the growing amount of heterogeneous data on the Web, it is important to make these data machine processable. In the Semantic Web, RDF data sets can be published with labels in different languages. In this context, data interlinking requires specific approaches to tackle cross-lingualism. We have evaluated two approaches based on machine translation and multilingual lexicon. Our results show that the best results are obtained using machine translation with an F-measure of 100%, while the results obtained with the multilingual lexicon are slightly lower with an F-measure of 88%. The highest results have been obtained on datasets with the rdfs:label property which shows that a name of a named entity is a discriminative feature in the interlinking process. Overall, both approaches seem to be promising for cross-lingual RDF data interlinking. However, the limitation would be the availability of language resources for a given pair of languages. The present work can be extended in the following directions:

- Test if both approaches can be complementary: errors made by one method can be corrected by the other method;
- Explore the suitability of Wikipedia for comparing resources.

ACKNOWLEDGMENTS

This work is partially supported by the ANR Lindicle[6] (12-IS02-0002) project in cooperation with Tsinghua University, China.

7. REFERENCES

[1] B. Fu, R. Brennan, and D. O'Sullivan. A Configurable Translation-Based Cross-Lingual Ontology Mapping System to adjust Mapping Outcome. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 15(3):15–36, 2012.

[2] J. Garcia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae. Challenges for the Multilingual Web of Data. *Journal of Web Semantics*, 11:63–71, 2012.

[3] T. Lesnikova, J. David, and J. Euzenat. Interlinking English and Chinese RDF Data Sets Using Machine Translation. In *Proceedings of the 3rd Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD 2014)*, volume 1243. CEUR-WS, 2014.

[4] C. Meilicke, R. García-Castro, F. Freitas, W. R. van Hage, E. Montiel-Ponsoda, R. R. de Azevedo, H. Stuckenschmidt, O. Svab-Zamazal, V. Svatek, A. Tamilin, C. Trojahn, and S. Wang. MultiFarm: A Benchmark for Multilingual Ontology Matching. *Journal of Web Semantics*, 15:62–68, 2012.

[5] R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.

[6] A. Nikolov, A. Ferrara, and F. Scharffe. Data linking for the semantic web. *Int. J. Semant. Web Inf. Syst.*, 7(3):46–76, July 2011.

[7] G. Sérasset and A. Tchechmedjiev. Dbnary: Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing, LREC 2014*, pages 68–71, 2014.

[8] Z. Wang, J. Li, Z. Wang, S. Li, M. Li, D. Zhang, Y. Shi, Y. Liu, P. Zhang, and J. Tang. XLore: A Large-scale English-Chinese Bilingual Knowledge Graph. In *Proceedings of the ISWC 2013 Posters & Demonstrations Track*, volume 1035, pages 121–124. CEUR-WS, 2013.

[6] http://lindicle.inrialpes.fr/