Interlinking English and Chinese RDF data sets using machine translation

Tatiana Lesnikova, Jérôme David, Jérôme Euzenat

Interlinking English and Chinese RDF Data Sets Using Machine Translation

Tatiana Lesnikova, Jérôme David, Jérôme Euzenat

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

Abstract. Data interlinking is a difficult task particularly in a multilingual environment like the Web. In this paper, we evaluate the suitability of a Machine Translation approach to interlink RDF resources described in English and Chinese languages. We represent resources as text documents, and a similarity between documents is taken for similarity between resources. Documents are represented as vectors using two weighting schemes, then cosine similarity is computed. The experiment demonstrates that TF*IDF with a minimum amount of preprocessing steps can bring high results.

Keywords: Semantic Web, Cross-Lingual Link Discovery, Cross-Lingual Instance Linking, owl:sameAs

1 Introduction

As there are resources (webpages) and links between them in the Web, so there are resources (real-world entities) and typed relationships between them in the Semantic Web. In the Semantic Web, several different URI references can refer to the same entity and the ability to identify equivalent entities is crucial for Linked Data. We assume that if two resources describe the same entity, these resources can be connected by means of an owl:sameAs link. The usage of owl:sameAs links has been studied in [1–3]. While there can be many relationships which describe different aspects of a resource (for example, "author", "journal", "date of publishing" for a scientific paper), we concentrate on owl:sameAs.

Interlinking resources scattered across heterogeneous data sources is not easy. This task can become particularly difficult due to the multilingual nature of the Web and information that can be found there. Apart from DBpedia with its multilingual versions [4] that became a central hub of the Linked Open Data (LOD), other publishers such as the French National Library [5], the Spanish National Library [6] make their data available using RDF model in their own language.

In this paper, we propose a method for interlinking RDF with multilingual labels and describe an experiment on interlinking resources with English and Chinese labels across two data sets. Given two RDF data sets, the goal is to find resources describing the same entity and set an owl:sameAs link between them.

The paper addresses the following questions:

- Can Machine Translation and the classical Information Retrieval (IR) vector-space model be suitable for interlinking RDF data?;
- How does the quality of generated owl:sameAs links depend on the data preprocessing techniques?

The rest of the paper is structured as follows. In the next section, we point to some recent works on multilingual resource interlinking; in Section 3 we describe our proposed method; Section 4 contains the description of the RDF data; Section 5 outlines evaluated parameters; Section 6 presents the evaluation of results and, finally, we draw conclusions in Section 7.

2 Related Work

The problem of searching for the same entity across multiple sources dates back to the 1960s. In database research, it is known as instance identification, record linkage or record matching problem. In [7], 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 related to ours, it does not contain cross-lingual aspect and RDF semantics.

In Natural Language Processing (NLP), the problems of entity resolution, multilingual entity recognition and cross-document co-reference resolution [8] gain a close attention due to their complexity and importance for Information Retrieval, Question Answering, etc. The task is to find out whether the occurrences of a name in different natural language texts refer to the same object. There is no general solution to this problem, and the decision whether two names refer to the same entity usually relies on contextual clues. Another related area is that of detecting the original text over its multilingual versions known as cross-lingual plagiarism detection [9].

In the Semantic Web, to facilitate data integration and knowledge sharing on the Web, interlinking tools capable of handling resources denoted in different natural languages are very important. 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 OAEI Data Interlinking track (IM@OAEI 2011), several interlinking systems have been proposed [10–12]. All of the systems were evaluated on monolingual data sets. Recent developments have been made also in multilingual ontology matching [13]. In [14], a systematic analysis was done to find the most effective string similarity metric for ontology alignment. An interesting aspect of this work is that it explores whether string preprocessing strategies such as tokenization, synonym lookup, translations, normalization, etc. can improve ontology alignment results. The authors mention that preprocessing procedures do not have a strong impact on performance, however they confirm the usefulness of Machine Translation (MT) when dealing with different languages. In contrast, we are not doing ontology matching in our case though we do use similarity metrics and data preprocessing.

The problem of instance-based interlinking of multilingual LOD has not been studied profoundly yet. The importance of cross-lingual mappings has been dis-

cussed in several works [15, 16]. For interlinking resources expressed in Asian languages, special methods for measuring string similarity are studied in [17, 18]. The work described in [19–21] shows the initiative of converting Chinese equivalent of Wikipedia (i.e. Baidu Baike and Hudong) into RDF data sets. The LIDER project facilitates multilingual, cross-media content analytics[1]. Some work has also been done in creating a multilingual ontology in RDF, e.g., BabelNet [22].

To the best of our knowledge, there is no interlinking system specifically designed to link RDF data sets with multilingual labels. In the next section, we sketch the principles of a method for this purpose.

3 Data Interlinking Method

We assume that the resources published in RDF are described in natural languages: property names and literals are usually natural language words. If so, then we hypothesize that NLP techniques can be used in order to detect the identical resources and interlink them. This means that our method is designed for RDF data sets which contain descriptions in natural language. It is inappropriate for RDF data sets containing purely numerical values.

The entire data flow with modifiable parameters is illustrated in Figure 1.

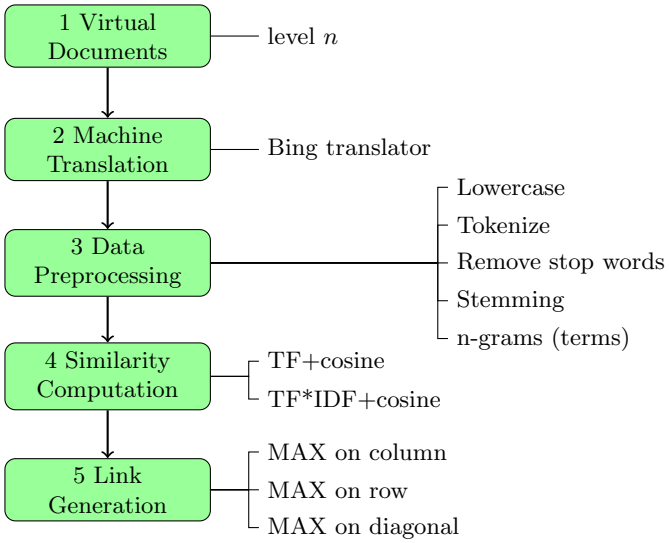

Fig. 1: Data Flow for Resource Interlinking

Given two RDF data sets, the method proceeds as follows.

[1] http://www.lider-project.eu/?q=what-is-lider

First, the resources are represented as **Virtual Documents** in different natural languages. The notion of a virtual document for RDF resources has already been described in [23, 24]. To obtain these virtual documents per resource, we collect literals according to the specified graph traversal distance, see Figure 2.

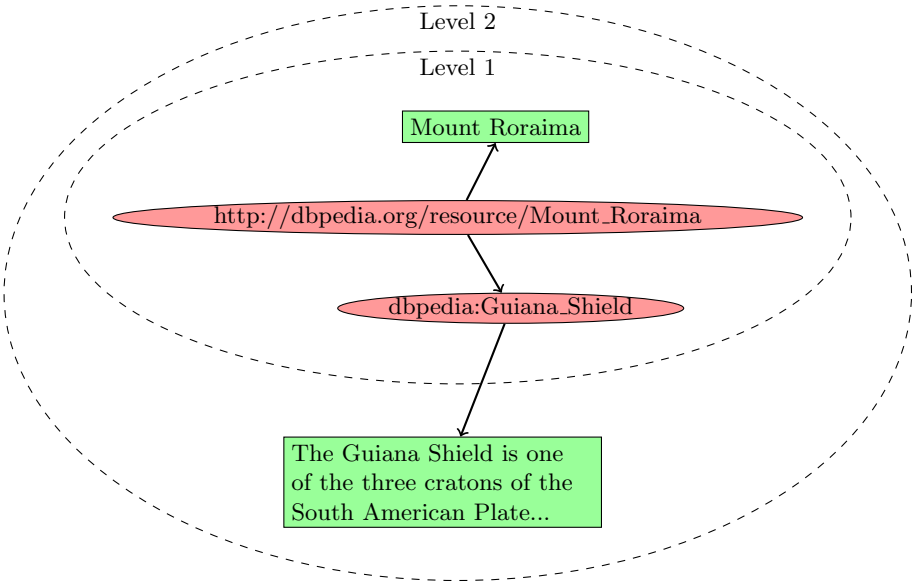

Fig. 2: Collecting Virtual Documents by Levels

Following the procedure above, we extract all the language information of a particular resource, for example, such properties as "label" and "comment" usually contain textual data. The purpose of this extraction is to form a virtual document which contains n levels of language information depending on the specified distance of graph traversal. The language elements attached to a particular type of relationships are taken into account. The property names are not considered. If the object is a literal, it is stored into a virtual document. If not, the algorithm proceeds to the following URI until it collects all the literals within a given distance.

Next, to make these documents comparable we use **Machine Translation**. Given two virtual documents in two different languages, it is important to make them comparable using a machine translation system. There are different kinds of MT: rule-based, statistical, hybrid. There are Google and Bing translator APIs. At this step, virtual documents in one language can be translated into the other language and vice versa or both languages can be translated into some third language.

Once translated, the documents undergo **Data preprocessing**. Translated virtual documents are treated as "bags of words", and different number of stan-

dard NLP preprocessing techniques (tokenization, stop word removal, etc.) are performed at this stage. We constructed four pipelines so that the number of processing steps is growing with each pipeline.

1. Pipeline 1 = Transform Cases into lower case + Tokenize;
2. Pipeline 2 = Pipeline 1 + Filter stop words;
3. Pipeline 3 = Pipeline 2 + Stem (Porter);
4. Pipeline 4 = Pipeline 3 + Generate n-grams (terms, max length = 2).

In order to compute similarity between the resources, we need to compute similarity between the documents that represent these resources. At **Similarity Computation** stage, various weighting schemes can be used for selecting the discriminant words. We chose two of them: Term Frequency (TF) and Term Frequency*Inverse Document Frequency (TF*IDF) and a similarity method to be applied, for example, the cosine similarity. The output of this stage is a similarity matrix. The matrix is such that the virtual documents in the original language are on the vertical axis and the translated documents are on the horizontal axis.

At **Link Generation** stage, the algorithm extracts links from the similarity matrix.

We study three ways of extracting links:

1. We select the maximum value in a column only (selecting the best original resource for a translation);
2. We select the maximum value in a row only (selecting the best translation for an original resource);
3. We select the maximum value in a column and a row (selecting such a translation for which the best original document has this translation as best translation).

4 Experimental Setup

Our goal is to evaluate how the method described above works and which parameters are important. We also evaluate the suitability of Machine Translation for identifying identical resources.

We would like to observe the effect of the size of virtual documents, preprocessing steps and weighting schemes (TF and TF*IDF) on the results. Basically, we seek an answer to the question: what is the combination of parameters that produces the highest results and can assure the correct match in the interlinking process?

4.1 Original RDF Data Sets

The experiment has been conducted on two separate RDF data sets with resources represented in English and Chinese natural languages respectively. Thus, the data consist of the English and Chinese part.

To fulfill the English part, we downloaded the following datasets from DBpedia 3.9[2]: Categories (Labels), Titles, Mapping-based Types, Mapping-based Properties, Short Abstracts, Extended Abstracts. For the Chinese part, we used a part of the Xlore.org[3] data: Abstracts, Reference Links to DBpedia, Inner Links, External Links, Infobox Property, Related Items, Synonyms. Xlore is the Chinese knowledge-base Baidu Baike converted into RDF.

All the data files have been accessed via a Jena Fuseki server and its built-in TDB store[4]. Statistics of data loaded into triple stores is presented in Table 1.

Table 1: Statistics about RDF Datasets

	# of classes	# of instances	# of properties	# of triples in total
DBpedia	435	3,220,000	1377	72,952,881
XLore	N/D	262,311	6280	7,063,975

4.2 Test RDF subset

We restricted our experiment to five entity types: Actors, Presidents, US Presidents, Sportsmen, and Geographical places. This was done for observing the difference in similarity within and across types.

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. Out of the reference link set provided by Xlore, we randomly selected 20 instances per category (Actors, Sportsmen, etc.) for which the two linked resources had text in their properties (more than just rdfs:label). In the US Presidents category, there were only 16 linked instances with text, this was compensated by adding four extra presidents into the category of Presidents.

This provided 100 pairs of entities potentially generating 10,000 links.

4.3 Protocol

The evaluation was carried out according to the following protocol:

- Provide the two sets of resources;
- Run a method configuration and collect the links;
- Evaluate links against the reference links through precision and recall.

5 Evaluated Configuration

The parameters evaluated are presented in Table 2. Thus, 48 settings have been explored in total.

[2] http://wiki.dbpedia.org/Downloads39
[3] http://xlore.org/index.action
[4] http://jena.apache.org/documentation/serving_data/

Table 2: Experimental parameters

VDocs 2	Pipelines 4	Translation 1	Weight 2	Similarity 1	Link Extraction 3
Level 1 Level 2	Pipeline 1 Pipeline 2 Pipeline 3 Pipeline 4	Bing: ZH→EN	TF TF*IDF	cosine	MAX on column MAX on row MAX on column and row

Translate ZH into EN

Once we collected a fixed number of entity pairs for each category in the English and Chinese data sets, we needed to make these entities comparable. For our experiment, we used the statistical translation engine: Bing Translator API[5] to translate Chinese virtual documents from the Chinese Simplified into the English language. Sometimes the large documents could not be translated in their entirety, in this case we left everything as is, taking only the part of text that has been translated. It would be interesting to translate documents from English into Chinese as well but our preprocessing tool does not support Asian languages, so at this point we were dealing only with translations from Chinese into English.

Data Preprocessing and Similarity Computation

The tool used for designing our pipelines was RapidMiner[6]. We were using RapidMiner 5.3.013 with the text processing extension.

Each data preprocessing step corresponds to a particular operator in RapidMiner. For some operators we can specify parameters. Below you can find the parameters used:

- Tokenize: mode: non-letters (i.e. non-letters serve as separators between tokens. Because of this, all dates are not preserved in documents);
- Filter Stopwords (English): built-in stopword list;
- The type of weighting scheme (TF or TF*IDF) was set for each pipeline;
- For computing similarity, we were using Data to Similarity Data operator with cosine similarity.

Link Generation

The output of the similarity computation is a matrix of compared pairs with a value. The 10,000 (100×100) comparisons were tabled as a similarity matrix for evaluation for each tested method. The matrix is such that the vertical axis represents the English DBpedia entities while the horizontal axis represents entities from the Chinese Xlore base.

[5] http://datamarket.azure.com/dataset/bing/microsofttranslator
[6] http://rapidminer.com/products/rapidminer-studio/

6 Results

The obtained results are displayed in Figures 3-4. They show that with TF*IDF/ Level 1 we are able to identify more than 97% of the identical entities. The comparison of virtual documents was done at two levels. The results across and within categories using TF*IDF show the same pattern: the best accuracy is achieved at Level 1 and the results get worse at Level 2. The results for TF were lower than those of TF*IDF so we do not report them here.

The similarity of resources within categories is presented in Figure 5. Black squares are 5 categories. The similarities are highlighted according to their value, and the color intensifies as the value grows:

- Values between 0.00 and 0.11 - are suppressed and seen as a white space;
- Values between 0.11 and 0.15 are in light yellow;
- Values between 0.15 and 0.25 are in dark yellow;
- Values between 0.25 and 0.35 are in orange;
- Values between 0.35 and 0.45 are in light red;
- Values between 0.45 and 1 are in dark red.

The correct match is always on the diagonal and the possible confusions are more likely within a category (see the last square (US Presidents)). This is expected since entities of the same type will have much information in common.

6.1 Discussion

The main points of the experiment are:

- Our results show the suitability of Machine Translation for interlinking multilingual resources;
- TF*IDF outperforms TF;
- The addition of preprocessing steps seem not to influence the results significantly. The maximum standard deviation is less than 2 points for both precision and recall;
- The quantity of information at Level 1 is usually enough to find a correct match;
- In general, the results at Level 2 were lower. This may be explained by supposing that the further we go from the node, the more general becomes the information. If there are many shared properties, then at some point many resources will have the same information (this can be due to the structure of the RDF data set). The discriminant information is thus "diluted" and it becomes harder to detect correct correspondences;
- If there is not enough data at Level 1 then by collecting information from Level 2 it is possible to improve the results. This gives us an intuition that the necessity of proceeding to the next level from Level 1 depends on the amount of data at Level 1. We saw this with one of the error cases when comparing across categories.

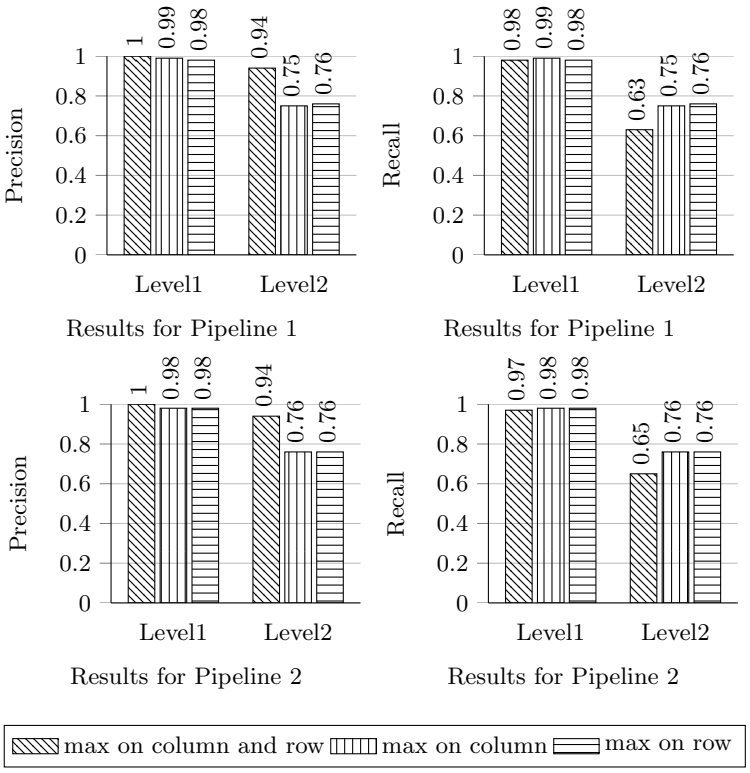

Fig. 3: Results for Level 1 and Level 2 using TF*IDF

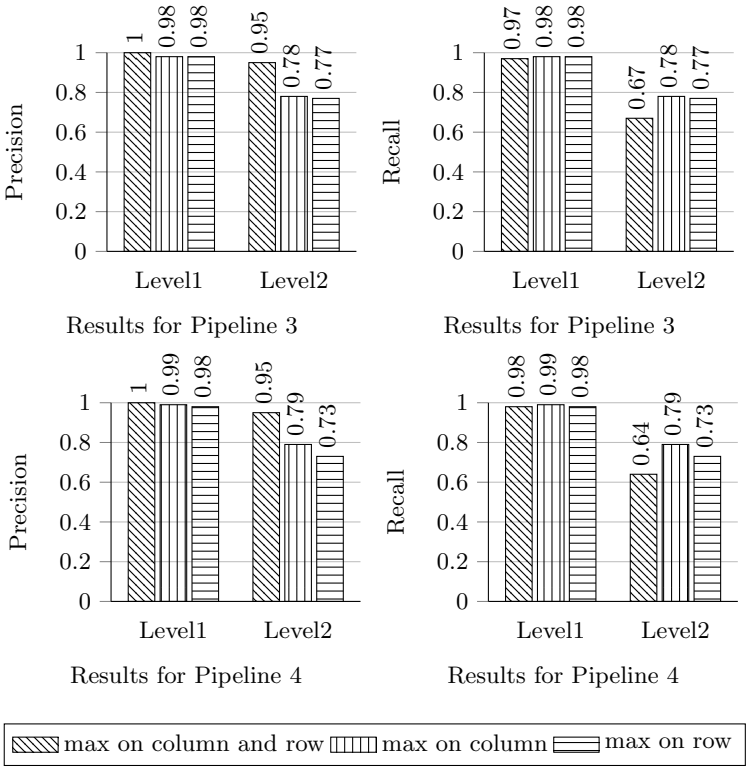

Fig. 4: Results for Level 1 and Level 2 using TF*IDF

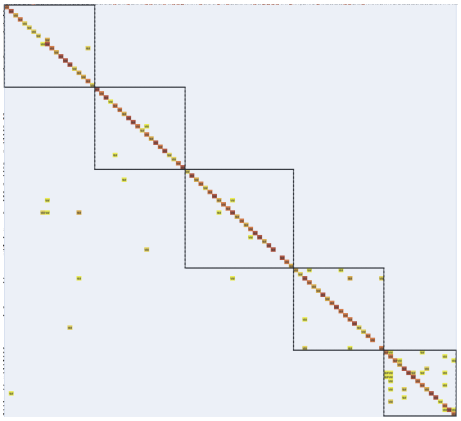

Fig. 5: Similarity within categories using TF*IDF at Level 1 Pipeline 1. Squares correspond to categories, and the darker the points, the higher the similarity. Dark points on the diagonal are correct matches. Most of the secondary dark points are confined in a square (a single category).

7 Conclusions and Future Work

Interlinking of resources described in different natural languages across heterogeneous data sources is an important and necessary task in the Semantic Web in order to enhance semantic interoperability. We described an instance-based interlinking method that mostly relies on labels and machine translation technology. The results demonstrated that the method can identify most of the correct matches using minimum information in a resource description with precision over 98%.

Though the reported results provide evidence that our method can be used for finding identical resources across two data sets, there are several axes that we currently left out of scope but will investigate in the future:

- Experimenting with other language pairs;
- Extending the coverage: adding other classes;
- Testing other similarity metrics;
- Exploiting other Machine Translation tools and evaluating their impact on the similarity computation;
- Exploring strategies that do not depend on translation technologies (e.g. mapping to WordNet).

Acknowledgments. This work has been done as part of the research within the Lindicle[7] (12-IS02-0002) project in cooperation with the Tsinghua University, China.

References

1. Jaffri, A., Glaser, H., Millard, I.: URI Disambiguation in the Context of Linked Data. In: Proc. of the WWW-2008 Workshop on Linked Data on the Web (LDOW-2008), Vol. 369, CEUR-WS.org, Beijing, China (2008)
2. Ding, L., Shinavier, J., Finin, T., McGuinness, D. L.: owl:sameAs and Linked Data: An Empirical Study. In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line (2010)
3. Community: Overloading OWL sameAs, `http://ontologydesignpatterns.org/wiki/Community:Overloading_OWL_sameAs`
4. Palmero Aprosio, A., Giuliano, C., Lavelli, A.: Towards an automatic creation of localized versions of DBpedia. In: Proc. of the 12th International Semantic Web Conference, ISWC 2013, Volume 8218, pp. 494–509 (2013)
5. 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)
6. 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)
7. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering. 19(1), 1–16 (2007)

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

8. 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, Vol.1, pp. 79-85 (1998)
9. Barrón-Cedeño Alberto, Gupta, P., and Rosso, P.: Methods for Cross-language Plagiarism Detection. J. Knowl.-Based Syst. 50, 211–217 (2013)
10. Volz, J., Bizez, C., Gaedke, M., and Kobilarov, G.: Discovering and Maintaining Links on the Web of Data. In: Proc. of ISWC' 09, pp. 650–665, Springer-Verlag Berlin, Heidelberg (2009)
11. 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)
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. 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)
14. Cheatham, M., Hitzler, P.: String Similarity Metrics for Ontology Alignment. In: Proc. of the 12th International Semantic Web Conference, Part II, LNCS, Vol. 8219, pp.294–309. Springer (2013)
15. 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)
16. 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)
17. Saemi Jang, Satria Hutomo, Soon Gill Hong, Mun Yong Yi: Interlinking Multilingual LOD Resources: A Study on Connecting Chinese, Japanese, and Korean Resources Using the Unihan Database. International Semantic Web Conference (Posters and Demos), pp. 229-232 (2013)
18. Soon Gill Hong, Saemi Jang, Young Ho Chung, Mun Yong Yi, Key-Sun Choi: Interlinking Korean Resources on the Web. JIST 2012: 382-387
19. 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)
20. Wang, Z., Li, J., Wang, Z., Tang, J.: Cross-lingual Knowledge Linking Across Wiki Knowledge Bases. In: Proc. WWW'12, pp. 459–468. ACM, NY (2012)
21. 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)
22. Navigli, R., Ponzetto, S.: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence. 193, 217-250 (2012)
23. 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. ACM Press (2006)
24. Lesnikova, T.: NLP for Interlinking Multilingual LOD. In: Proc. of the Doctoral Consortium at the 12th International Semantic Web Conference (ISWC 2013), Vol.1045 of CEUR Workshop Proceedings, pp. 32–39. CEUR-WS.org (2013)