Cross-lingual RDF Thesauri Interlinking

Tatiana Lesnikova, Jérôme David, Jérôme Euzenat
INRIA & University of Grenoble Alpes, Grenoble, France
tatiana.lesnikova@inria.fr, jerome.david@inria.fr, jerome.euzenat@inria.fr

Abstract

Various lexical resources are being published in RDF. To enhance the usability of these resources, identical resources in different data sets should be linked. If lexical resources are described in different natural languages, then techniques to deal with multilinguality are required for interlinking. In this paper, we evaluate machine translation for interlinking concepts, i.e., generic entities named with a common noun or term. In our previous work, the evaluated method has been applied on named entities. We conduct two experiments involving different thesauri in different languages. The first experiment involves concepts from the TheSoz multilingual thesaurus in three languages: English, French and German. The second experiment involves concepts from the EuroVoc and AGROVOC thesauri in English and Chinese respectively. Our results demonstrate that machine translation can be beneficial for cross-lingual thesauri interlining independently of a dataset structure.

Keywords: Cross-lingual Data Interlinking, owl:sameAs, Thesaurus Alignment

1. Introduction

In the Semantic Web, entities are described in triples (subject, predicate, object) following the W3C Resource Description Framework (RDF) (Lassila and Swick, 1999). Most commonly, entities are real-world individuals and events. Moreover, linguistic resources such as thesauri and dictionaries are also available in RDF. These linguistic resources should be interlinked to enhance their interoperability (McCrae et al., 2012).

In this paper, we evaluate a translation-based interlinking method on terminology expressed in different natural languages. This interlinking method has been applied to the encyclopedic resources from DBpedia (Bizer et al., 2009) in English and XLore (Wang et al., 2013) in Chinese on which we obtained good results (Lesnikova et al., 2014). Though this method has been initially developed for cross-lingual interlinking RDF instances, we consider its application to linking heterogeneous multilingual *linguistic* resources. There are many lexical-semantic resources for different languages and domains grouped in the Linguistic Linked Open Data cloud[1] (Chiarcos et al., 2011).

We address the following problem: Given two thesauri with labels in two different languages, find equivalent concepts and link them using owl:sameAs relation.

We represent concepts as text documents containing labels in a respective language, documents are translated and represented in a vector space model. Similarity between documents is taken for similarity between concepts. Extraction of matches is based on the similarity values. A set of RDF statements forms a labeled directed graph where nodes represent concepts and edges represent relationships between these concepts. An RDF thesaurus is a graph where resources are concepts labeled in natural languages. A context of a concept are the labels of the neighboring nodes. However, a context can be very narrow if there is not much textual information in the concept description.

The remainder of the paper is structured as follows. Section 2 presents related work on multilingual lexical resources and methods for their interlinking. Section 3 briefly presents the approach for interlinking multilingual entities. Section 4 describes test data sets used in the experiments and evaluation parameters. The results of the experiments are discussed in Section 5. Section 6 draws conclusions and points to future work.

2. Related Work

We identify two areas of related work: a) the development of multilingual vocabularies; b) the development of methods for interlinking such resources in order to enhance the data exchange between the systems which use these resources.

2.1. Multilingual Vocabularies

The notion of a knowledge organization system has been developed in library and information sciences. Such a system organizes information by means of controlled vocabularies such as classification schemes, subject heading, taxonomies and thesauri. Thesauri consist of a predefined list of terms or short phrases aimed at cataloging information to facilitate its retrieval. Thesauri contain concepts and the relationships between them. The relationships between concepts usually include hierarchy, synonymy and relatedness (Hodge, 2000).

There are various thesauri published as linked data and thus available in a machine-readable format on the Web. SKOS (Simple Knowledge Organization System) (Miles and Bechhofer, 2009) is an ontology widely used for representing conceptual hierarchies on the Web. The Environmental Applications Reference Thesaurus (EARTh) (Albertoni et al., 2014) is a SKOS multilingual dataset containing terms related to the environment. Other thesauri available as Linked Data are GEneral Multilingual Environmental Thesaurus (GEMET)[2], AGROVOC[3], EUNIS[4], Geologi-

[1] http://linguistics.okfn.org/resources/llod/

[2] http://www.eionet.europa.eu/gemet/en/themes/

[3] http://aims.fao.org/standards/agrovoc/concept-scheme

[4] http://datahub.io/dataset/eunis

cal Survey of Austria (GBA) Thesaurus[5] - a bilingual (German/English) vocabulary for representing geodata. These terminological resources are interlinked. For example, EARTh thesaurus has links to GEMET, AGROVOC as well as DBpedia. EuroVoc[6] is a multilingual thesaurus in 23 languages covering a wide range of activities of the European Union.

Lexical semantic resources are also published as linked data: WordNet (van Assem et al., 2006); BabelNet (Navigli and Ponzetto, 2012) – a multilingual dictionary which covers 271 languages in BabelNet 3.0 edition; DBnary (Sérasset and Tchechmedjiev, 2014) contains extracted data from Wiktionary[7] in 12 languages. The LIDER project[8] aims at providing interlinked language resources (corpora, dictionaries, etc.) for exploitation in multilingual content analytics across different media resources.

2.2. Cross-lingual RDF Data Interlinking

The problem of searching for the same entity originated in database research, and it is known as instance identification, record linkage or record matching problem. A thorough survey on the matching techniques is provided in (Elmagarmid et al., 2007). The present work is related to the record linkage in the sense that the duplicate concepts should be detected, however the search for duplicate records is done within a single data source complying to the same schema and it contains neither cross-lingual aspect nor RDF semantics or ontologies.

In Natural Language Processing (NLP), the task of identifying whether the occurrences of a name in different natural language texts refer to the same object is known as entity resolution and cross-document co-reference resolution (Bagga and Baldwin, 1998). Another related area is that of detecting the original text over its multilingual versions known as cross-lingual plagiarism detection (Barrón-Cedeño et al., 2013).

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. Data interlinking (Ferrara et al., 2011) is the process of setting sameAs links between semantically related entities, i.e., entities referring to the same object. Cross-lingual interlinking consists in discovering links between identical resources across diverse RDF sources in different languages. It is one of the challenges for the multilingual Web of data (Gracia et al., 2012). Within the Ontology Alignment Evaluation Initiative Data Interlinking track (IM@OAEI), most systems are evaluated on monolingual data (Ngonga Ngomo and Auer, 2011; Araújo et al., 2011).

Recent developments have been made in multilingual ontology matching (Meilicke et al., 2012). The distinction between multilingual matching and cross-lingual matching is considered in (Spohr et al., 2011; Euzenat and Shvaiko, 2013). A common approach to bridge the natural language gap is to transform a cross-lingual problem into a monolingual one by translating the elements of one ontology into the language of the other ontology (Fu et al., 2012) using machine translation. After translation, monolingual matching strategies (Euzenat and Shvaiko, 2013) are applied. In (Wang et al., 2009; Fu et al., 2009; Trojahn et al., 2010), the Google Translate API service has been used. In the MultiFarm track[9] of OAEI 2014, the AML system achieved the highest F-measure of 0.54 using Microsoft Bing Translator. Another way to approach ontology matching is to use external lexical resources. Some of the ontology matching approaches employ Wikipedia's search functionality and interlanguage links for finding mappings (Hertling and Paulheim, 2012). In (Lin and Krizhanovsky, 2011), Wiktionary[10] is used as a lexical background knowledge.

If two ontologies contain concepts with multiple labels in overlapping languages, this multilingual information can be very useful for matching (Spohr et al., 2011). Multilingual techniques which take advantage of multiple labels for finding correspondences between concepts from EuroVoc-AGROVOC thesauri are evaluated in (Dragoni, 2015). Simple string matching techniques on the overlapping languages have been used to link AGROVOC to other thesauri (Morshed et al., 2011). The use of string matching techniques may be impractical if two languages use different alphabets.

In (Lesnikova et al., 2014), we proposed an interlinking method and evaluated it on resources described in the English and Chinese languages. We translated terms of one dataset into the language of the other dataset, i.e., the entity linking was done cross-lingually. The limitations of evaluation lied in the small size of the corpus and the nature of the resources: all resources represented Named Entities, e.g., actors, geographical places, etc. However, we consider our method applicable to any type of resources.

The main aspects in which our present work is different from the previous one are:

1. We use a linguistic resource (a thesaurus) instead of a classical linked dataset such as DBpedia;
2. The nature of instances to be linked is different: instead of concrete entities we use general concepts;
3. We evaluate two different methods for alignment extraction.

In this work, we assume that two RDF data sets contain labels in two different languages, so multilingual matching is not suitable. Also, we investigate the importance of a concept context in order to establish a correct match. In the next section, we outline the main components of the method for this purpose.

3. Translation-based Interlinking Method

The interlinking method consists of five steps:

1. Constructing a **Virtual Document** in different languages per resource.

[5] http://datahub.io/dataset/geological-survey-of-austria-thesaurus

[6] http://eurovoc.europa.eu/

[7] https://www.wiktionary.org/

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

[9] http://oaei.ontologymatching.org/2014/multifarm/results/index.html

[10] www.wiktionary.org

2. Translating documents using **Machine Translation** in order to transform documents into the same language. At this step, virtual documents in one language are translated into the other language and vice versa or both languages can be translated into some pivot language.

3. Cleaning documents using **Data preprocessing** techniques such as tokenization, stop-word removal, etc. We use the following text preprocessing: Transform Cases into lower case + Tokenize + Filter stop words.

4. **Computing Similarity** between documents using term weights and applying similarity methods, for example, the cosine similarity. The output of this step is a set of similarity values between pairs of virtual documents.

5. **Generating Links** between concepts. The goal of interlinking is to identify a set of correspondences between concepts. At this stage, an algorithm extracts links on the basis of the similarity between documents. There are different methods to extract matches (Euzenat and Shvaiko, 2013).

Virtual Documents are constructed as follows.

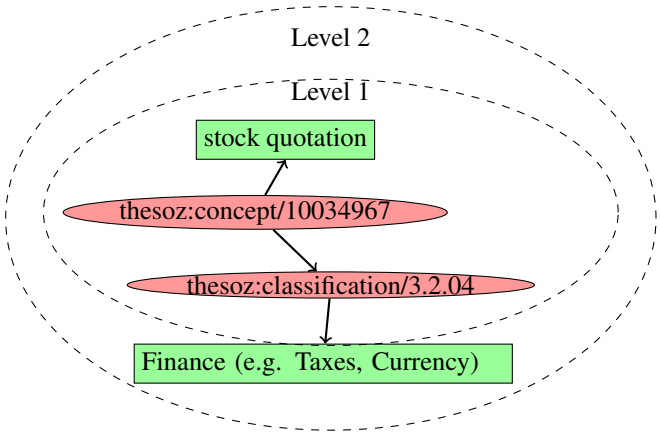


Figure 1: Creation of Virtual Documents by Levels

Due to the graph structure of RDF, we can collect literals according to the specified graph traversal distance (Level 1, Level 2, etc.), see Figure 1. The triples of an RDF graph can have simple strings (literals) as an object which serve as a descriptor for a subject. In the example of Figure 1, the subject is "thesoz:concept/10034967" which has a label "stock quotation". These collected literals will constitute the body of the virtual documents. The performance of the method may depend on the amount of text and discriminative power of labels. At this step, we also suppress all metadata information about the dataset: for example, objects of "http://purl.org/dc/terms/" property describe creators of the dataset, dates of creation and modification. The properties to remove were detected by observing the generated documents. Thus, a virtual document contains only proper lexical items, the names of the properties themselves are also omitted.

An example of a virtual document at Level 1 before suppressing metadata:

```
stock quotation
4.6.07
```

An example of a virtual document at Level 1 after suppressing metadata:

```
stock quotation
```

An example of a virtual document at Level 2 before suppressing metadata:

```
Descriptors of the TheSoz
2014-08-14
GESIS - Leibniz Institute for the
                Social Sciences
...
stock quotation
Finance (e.g. Taxes, Currency)
4.6.07
```

An example of a virtual document at Level 2 after suppressing metadata:

```
stock quotation
stock quotation
Finance (eg Taxes, Currency)
```

4. Evaluation Setup

We conduct two experiments. The first experiment evaluates machine translation on concepts from the TheSoz multilingual thesaurus in three languages: English, French and German. Even though the obtained results are high, it might be due to the same structure of concept descriptions as the concepts belong to the same thesaurus. To verify that machine translation results are independent of the knowledge structure, we conducted another experiment involving two different thesauri. The second experiment evaluates machine translation on concepts in English and Chinese from EuroVoc and AGROVOC respectively.

4.1. Data

For the first experiment, we use the multilingual thesaurus for the Social Sciences - TheSoz 0.93 (Zapilko et al., 2013). This is a SKOS-based thesaurus containing concepts with labels in English, German and French languages. There are 8223 concepts in total for each language. 12 of them have no English label, and 6 concepts do not have French label. In the experiments, we use the 8206 concepts shared by the three languages. In order to provide a reference alignment, we split the thesaurus into three language specific datasets which contain the same concepts with a label in a respective language. Since the same URI identifies a given concept in each language, we could compare the obtained links against the reference. The reference contains 8206 links in which concepts are in one-to-one correspondence.

For the second experiment, we use multilingual thesauri from multidisciplinary and agricultural domains: EuroVoc and AGROVOC. We extracted entities from the existing reference alignment (1318 entities linked by "skos:exactMatch" property). We suppressed duplicate concepts from EuroVoc and their corresponding concepts from AGROVOC. In the experiments, we use the 1300 concepts in English and Chinese. The reference contains 1300 links in which concepts are in one-to-one correspondence.

In the evaluated experiments, we use only one pair of languages at a time, i.e., English vs. French, English vs. Chinese, German vs. French, etc.

4.2. Evaluated Configuration

The parameters evaluated are presented in Figure 2.

Virtual Documents. For the first experiment, we constructed virtual documents for Level 1 and Level 2 for the three language pairs. After the results were obtained, we decided to build virtual documents at Level 3 for the best language pair in order to see whether a larger context affects the results. For the second experiment, we built virtual documents for 3 levels.

Translation of non-English virtual documents into English. All source languages have been translated into English. We used the statistical translation engine Bing Translator[11] to perform all translations.

For the experiment with the TheSoz thesaurus, 2 types of comparison should be noted:

- **French and German translation to English.** The collected virtual documents from the English and French/German data sets are made comparable by translating French and German into one common language (English). Thus, if the French virtual documents are compared with the German ones, English is a pivot language.
- **German translation to French.** In order to verify that the way the virtual documents are translated can affect the results, we also translate German into French, and compare the translated documents against the original French dataset. In this case, the translation is done directly from the source language (DE) into the target one (FR).

Data Preprocessing and Similarity Computation. RapidMiner[12] 5.3.013 with the text processing extension was used for document preprocessing. Each data preprocessing step corresponds to a particular operator in RapidMiner. The following configurations were used:

- Tokenize: mode: non-letters;
- Filter Stopwords (English, French): built-in stopword lists;
- The TF*IDF weighting scheme was used in all settings;
- For computing similarity, we were using Data to Similarity Data operator with cosine similarity.

Link Generation. The output of the similarity computation is a set of similarity values between compared entity pairs. We use the Hungarian (Munkres, 1957) and greedy algorithms to extract the matches. The Hungarian algorithm yields the global optimum while the greedy algorithm yields a local optimum. We suppressed null similarities for match extraction.

Randomly removed concepts. The original 8206 concepts common to three language-specific datasets are in a one-to-one relationship with each other. We conducted an additional experiment in order to see how the similarity behaves if concepts in one dataset do not appear in the other one. This experiment has been done on the language pair which showed the highest results using the evaluated configuration in section 4.2: EN-DE language pair. We randomly suppressed 40% of concepts from both datasets and only 60% of the concepts has been preserved. Thus, out of 8206 original concepts, only 4943 concepts took part in the experiment, 2995 out of which constituted reference links.

4.3. Protocol

The evaluation was carried out according to the following protocol:

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

5. Results

The interlinking results of concepts from the TheSoz thesaurus are presented in Section 5.1. The interlinking results of concepts from the EuroVoc-AGROVOC are presented in Section 5.2. Each subfigure shows results for a particular language pair using both link extraction algorithms, we compute the F-measure for each setting and present it on the y-axis. The legend is the same for all figures as for Figure 3.

5.1. Linking concepts from TheSoz

Figure 3, Figure 4 and Figure 5 demonstrate that the F-measure grows with level. The best F-measure of 0.91 was found at Level 3 which is an improvement of 26 percentage points compared to Level 1.

The results using English as pivot language are better than direct translation between German and French. The results where French and German virtual documents have been translated into English and compared against the original English data are provided in Figure 3. The results of comparison against French original data where German virtual documents have been translated into French are presented in Figure 4.

The evaluated method seems to be not very robust when comparisons between concepts are not one-to-one. Figure 5 shows the results of the additional experiment with randomly removed concepts. The best matches are obtained at Level 2 and 3 with F-measure of 0.59 for the Hungarian method.

Concerning the link extraction methods, both link extraction algorithms obtained relatively similar results at Level 1. The Hungarian algorithm outperformed the greedy one at Level 2 and Level 3 and showed an increase of F-measure.

In the present experiment, the obtained results are different from results obtained with Named Entities. In previous experiments (Lesnikova et al., 2014), the cross-lingual interlinking has been done between resources representing Named Entities, and the method could identify most of the correct matches with the F-measure over 0.95 at Level 1.

The quantity of information in virtual documents can influence the output of machine translation. Level 1 often contains a single word or a short phrase. If machine translation

[11] http://datamarket.azure.com/dataset/bing/microsofttranslator

[12] http://rapidminer.com/products/rapidminer-studio/

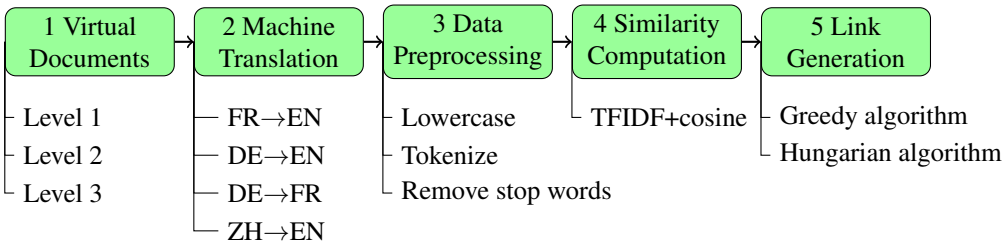
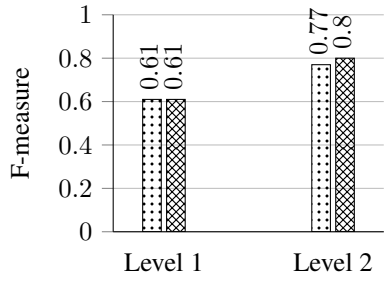
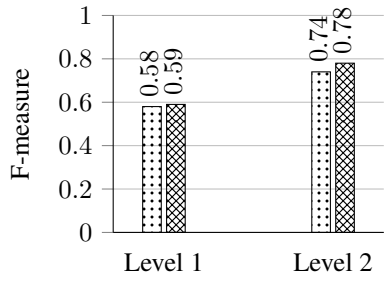


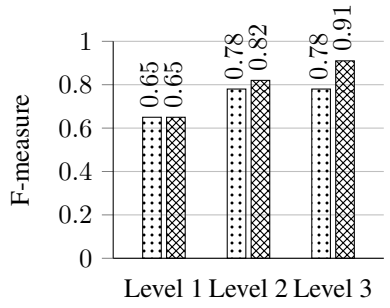
Figure 2: Experimental parameters



(a) Results for the EN-FR language



(b) Results for the FR-DE language



(c) Results for the EN-DE language pair

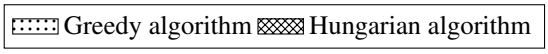


Figure 3: French and German languages are translated into English and compared against the English original data. For FR-DE pair, English is a pivot language. Results for Level 1, Level 2 and Level 3 using TF·IDF.

is not exact at Level 1, the mismatch is possible. That is why it is important to extend the context of a term by proceeding to further levels.

The best results are obtained for the English-German language pair (Figure 3). The worst results relate to the French-German language pair when the German language has been directly translated into French (Figure 4).

The results of the experiment with randomly removed concepts (Figure 5) show again that the similarity between en-

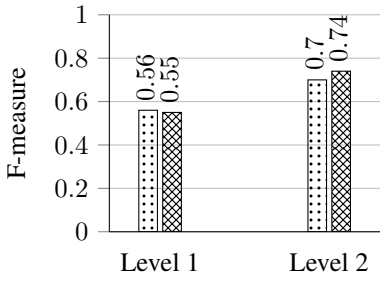


Figure 4: Results for the FR-DE language pair. German language is translated into French, comparison done against French original data.

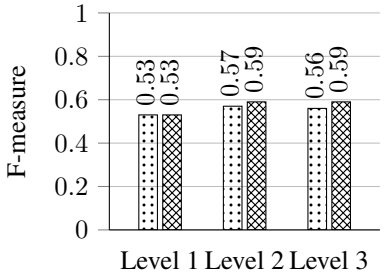


Figure 5: Results for the EN-DE language pair. 40% of the concepts have been randomly removed from both datasets.

ties grows as the level increases: precision has been relatively the same across all levels, and we observed an increase of recall by at least 10 percentage points from level 1 to further levels. Though the results are lower, the correct matches have got the highest similarity values even when resources are not in a one-to-one relationship.

The conducted evaluation showed a different performance of the interlinking method when tested on the resources represented by generic terms (a concept label is usually a common noun or a term in a thesaurus). Thus, it seems that it is more difficult to interlink concepts of a thesauri rather than resources corresponding to named entities.

5.2. Linking concepts from EuroVoc-AGROVOC

The results where Chinese virtual documents have been translated into English and compared against the original English data are provided in Figure 6. The main difference with the TheSoz results is that F-measure drops as levels grow. The best F-measures of 0.81 and 0.80 at Level 1 were

obtained by both link extraction algorithms[13]. The results at Level 3 dropped significantly (by 20 percentage points) for both algorithms. The decrease of the results at Level 3 can be due to the difference in knowledge organization of each thesaurus.

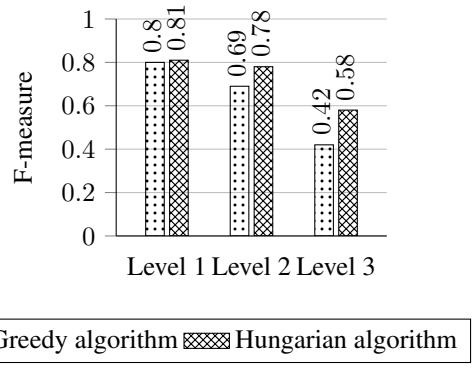


Figure 6: Results on concepts from EuroVoc-AGROVOC on the EN-ZH language pair.

5.3. Comparison of Results According to a Threshold

The best results of both link extraction algorithms are evaluated according to a threshold. Figures 7 and 8 present the best results for the TheSoz concept linking, i.e., for the English-German language pair according to the results in Figure 3. Figures 9 and 10 present the results for the EuroVoc-AGROVOC concept linking. The threshold corresponds to a similarity value for which extracted links were evaluated. The purpose of this evaluation was to observe if the results change drastically after a certain threshold. We could observe that the F-measure decreases in all cases because recall decreases faster than precision increases. Overall, the correct matches are distributed across a wide range of similarity values, so establishing the threshold above 0 may not provide the best cutoff.

6. Conclusions

This paper evaluated machine translation on interlinking terminology expressed in different natural languages. We observed the impact of the quantity of textual information in resource description by collecting information from further removed neighboring nodes. We evaluated the approach on 8206 concepts in English, French and German languages from the TheSoz thesaurus. The best F-measure of 0.91 has been obtained at Level 3 on the EN-DE language pair by the Hungarian algorithm. We also evaluated this method on the concepts in English and Chinese from the EuroVoc and AGROVOC thesauri. The best F-measure of 0.81 has been obtained at Level 1 by the Hungarian algorithm. The results of both experiments demonstrated that machine translation can work well independently of a dataset structure. The present evaluation shows that the similarity-based method can be applied on resources which do not necessarily contain a named entity as their label, though it is harder to find a correct correspondence in this case. The directions for future work may include: (1) context-based matching: matching the French-German DBpedia through the TheSoz mediation and the French-Chinese DBpedia-XLore matching through XLore mediation; (2) using external lexical resources for cross-lingual data interlinking.

[13] An F-measure of 0.82 is reported in (Dragoni, 2015). However, the number of reference links reported is different.

Acknowledgments

We thank Benjamin Zapilko from GESIS for providing us with language specific versions of the TheSoz. This work has been done as part of the research within the Lindicle[14] ANR (12-IS02-0002) project.

Bibliographical References

Albertoni, R., Martino, M. D., Franco, S. D., Santis, V. D., and Plini, P. (2014). EARTh: An Environmental Application Reference Thesaurus in the Linked Open Data cloud. *Semantic Web journal (SWJ)*, 5:165–171.

Araújo, S., Hidders, J., Schwabe, D., and de Vries, A. P. (2011). SERIMI - Resource Description Similarity, RDF Instance Matching and Interlinking. *CoRR*, abs/1107.1104.

Bagga, A. and Baldwin, B. (1998). 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*, volume 1, pages 79–85.

Barrón-Cedeño, A., Gupta, P., and Rosso, P. (2013). Methods for cross-language plagiarism detection. *J. Knowl.-Based Syst.*, 50:211–217.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3):154–165.

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2011). Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, 52(3):245–275.

Dragoni, M. (2015). Exploiting multilinguality for creating mappings between thesauri. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC'15, pages 382–387, New York, NY, USA. ACM.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16.

Euzenat, J. and Shvaiko, P. (2013). *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition.

Ferrara, A., Nikolov, A., and Scharffe, F. (2011). Data linking for the semantic web. *International Journal of Semantic Web and Information Systems*, 7(3):46–76.

Fu, B., Brennan, R., and O'Sullivan, D. (2009). Cross-Lingual Ontology Mapping — An Investigation of the Impact of Machine Translation. In *Proceedings of the 4th Asian Conference on The Semantic Web*, ASWC '09, pages 1–15. Springer-Verlag.

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

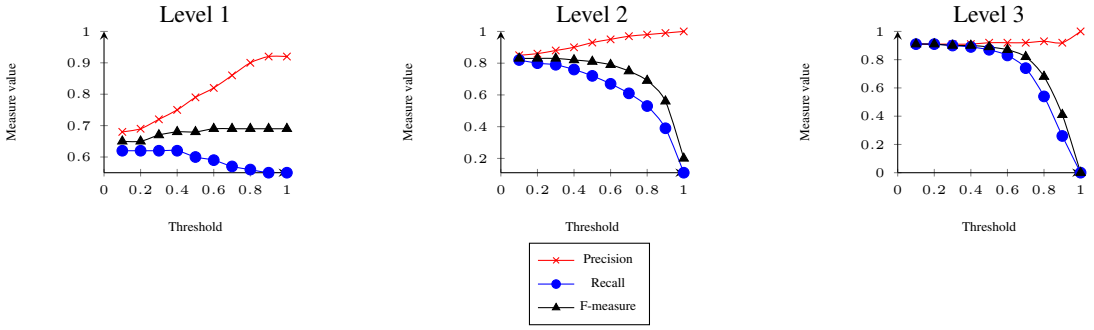


Figure 7: Hungarian results for the TheSoz: EN-DE language pair.

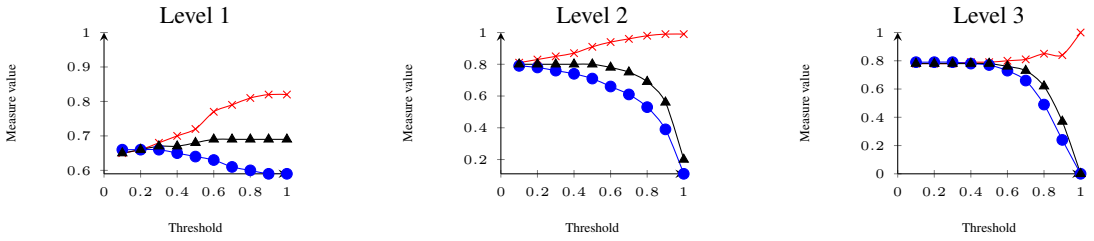


Figure 8: Greedy results for the TheSoz: EN-DE language pair.

Fu, B., Brennan, R., and O'Sullivan, D. (2012). 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).

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., and McCrae, J. (2012). Challenges for the Multilingual Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:63–71, March.

Hertling, S. and Paulheim, H. (2012). WikiMatch - Using Wikipedia for Ontology Matching. In *Proceedings of the 7th International Workshop on Ontology Matching*, volume 946 of *CEUR Workshop Proceedings*, Boston, MA, USA, November. CEUR-WS.org.

Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Technical report, Council on Library and Information Resources, Washington, DC. Digital Library Federation.

Lassila, O. and Swick, R. R. (1999). Resource Description Framework (RDF) Model and Syntax Specification. Technical report, World Wide Web Consortium.

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

Lin, F. and Krizhanovsky, A. (2011). Multilingual Ontology Matching Based on Wiktionary Data Accessible via SPARQL Endpoint. In *Proceedings of the 13th Russian Conference on Digital Libraries*, RCDL'2011, pages 19–26, Voronezh, Russia, October 19-22.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. In *Language Resources and Evaluation*, volume 46(4), pages 701–719. Springer.

Meilicke, C., GarcÃa-Castro, R., Freitas, F., van Hage, W. R., Montiel-Ponsoda, E., de Azevedo, R. R., Stuckenschmidt, H., Å vÃ¡b Zamazal, O., SvÃ¡tek, V., Tamilin, A., Trojahn, C., and Wang, S. (2012). MultiFarm: A Benchmark for Multilingual Ontology Matching. *Journal of Web Semantics*, 15:62–68.

Miles, A. and Bechhofer, S. (2009). SKOS Simple Knowledge Organization System Reference.

Morshed, A., Caracciolo, C., Johannsen, G., and Keizer, J. (2011). Thesaurus Alignment for Linked Data publishing. In *International Conference on Dublin Core and Metadata Applications*, pages 37–46.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Ngonga Ngomo, A.-C. and Auer, S. (2011). LIMES: A time-efficient approach for large-scale link discovery on the web of data. In *Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2312–2317. AAAI Press.

Sérasset, G. and Tchechmedjiev, A. (2014). 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.

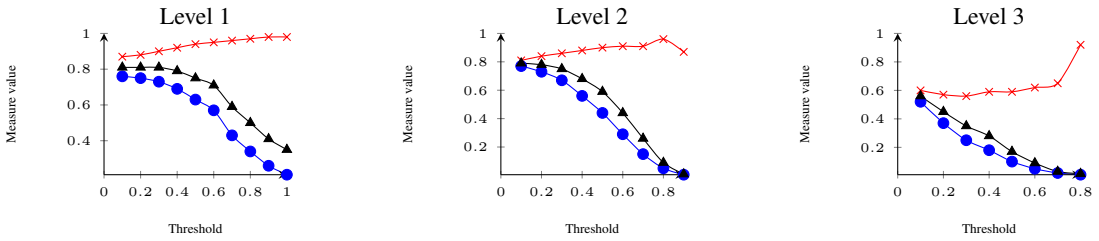


Figure 9: Hungarian results for the EuroVoc-AGROVOC: EN-ZH language pair.

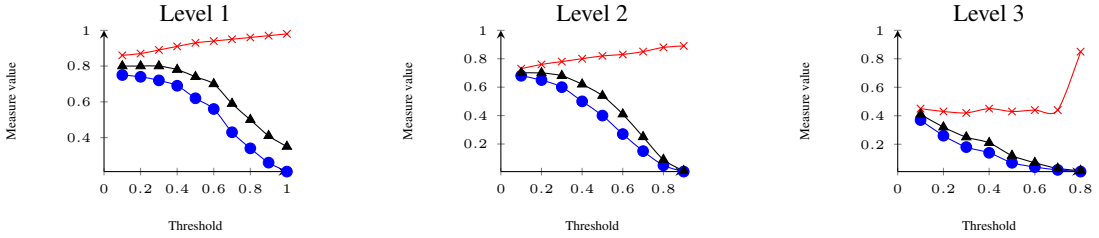


Figure 10: Greedy results for the EuroVoc-AGROVOC: EN-ZH language pair.

Spohr, D., Hollink, L., and Cimiano, P. (2011). A Machine Learning Approach to Multilingual and Cross-lingual Ontology Matching. In *Proceedings of the 10th International Conference on The Semantic Web - Volume Part I*, ISWC'11, pages 665–680. Springer-Verlag.

Trojahn, C., Quaresma, P., and Vieira, R. (2010). An API for Multi-lingual Ontology Matching. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3830–3835, Valletta, Malta. European Language Resources Association (ELRA).

van Assem, M., Gangemi, A., and Schreiber, G. (2006). Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the LREC (2006)*, pages 237–242.

Wang, S., Isaac, A., Schopman, B. A. C., Schlobach, S., and van der Meij, L. (2009). Matching Multilingual Subject Vocabularies. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, volume 5714, pages 125–137. Springer, Heidelberg.

Wang, Z., Li, J., Wang, Z., Li, S., Li, M., Zhang, D., Shi, Y., Liu, Y., Zhang, P., and Tang, J. (2013). XLore: A Large-scale English-Chinese Bilingual Knowledge Graph. In *International Semantic Web Conference (Posters & Demos)*, volume 1035 of *CEUR Workshop Proceeding*, pages 121–124. CEUR-WS.org.

Zapilko, B., Schaible, J., Mayr, P., and Mathiak, B. (2013). TheSoz: A SKOS Representation of the Thesaurus for the Social Sciences. *Semantic Web journal (SWJ)*, 4(3):257–263.