Selecting Optimal Background Knowledge Sources for the Ontology Matching Task

Abdel Nasser Tigrine, Zohra Bellahsene, Konstantin Todorov

Selecting Optimal Background Knowledge Sources for the Ontology Matching Task

Abdel Nasser Tigrine, Zohra Bellahsene, Konstantin Todorov

{lastname}@lirmm.fr
LIRMM / University of Montpellier, France

Abstract. It is a common practice to rely on background knowledge (BK) in order to assist and improve the ontology matching process. The choice of an appropriate source of background knowledge for a given matching task, however, remains a vastly unexplored question. In the current paper, we propose an automatic BK selection approach that does not depend on an initial direct matching, can handle multilingualism and is domain independent. The approach is based on the construction of an index for a set of BK candidates. The couple of ontologies to be aligned is modeled as a query with respect to the indexed BK sources and the best candidate is selected by following an information retrieval paradigm. We evaluate our system in a series of experiments in both general-purpose and domain-specific matching scenarios. The results show that our approach is capable of selecting the BK that provides the best alignment quality with respect to a given reference alignment for each of the considered matching tasks.

1 Introduction

Over the past years, the web has been continuously evolving from a web of documents to a web of data, following the principles of data and knowledge representation, publishing and linking. The Linked Open Data project[1], the web of data most successful initiative to date, comprises nowadays hundreds of datasets over several domains of life and science. While information is expressed by the help of RDF (Resource Description Framework) statements, knowledge about the domains of interest is given in the form of ontologies, which provide common vocabularies to name classes of things (concepts) and relations between these classes, defining in an explicit manner their semantics. Ontologies, expressed in RDFS or OWL, can be simple sets of terms, thesauri or more complex structured vocabularies with logical expressions that allow for the inference of new facts.

It occurs often that ontologies, describing similar or equivalent domains of knowledge, are expressed differently. These differences, referred to as ontology heterogeneities, can occur in terms of terminology (choosing different names to refer to the same concepts and relations), structure or semantics (relating classes in different ways, giving different intensions to information) or simply in terms of

[1] `http://linkeddata.org`

syntax (choosing different formal representations). In order to unlock the potential of the web of data and foster the creation of a veritable information network, heterogeneous ontologies have to be linked together by explicitly declaring the equivalence relations between their entities (classes and properties). The field of ontology matching has taken the challenge of proposing solutions that allow to automatically discover the correspondence between ontological elements in the presence of one or more of the heterogeneities cited above. As a result of almost 20 years of research and practice, many approaches and systems exist, capable of aligning highly heterogeneous ontologies [1].

It has been shown in recent publications [2,3] that the ontology matching process can benefit largely from the use of Background Knowledge (BK). BK is understood as any external reference knowledge that can facilitate the matching process, given in the form of large general purpose ontologies or well-established knowledge graphs (such as DBPedia or YAGO), domain specific ontologies, or the web at large. We outline three main advantages of the use of BK when aligning ontologies. In the first place, as observed by [4], there is always an inherent semantic gap between two ontologies, coming from the missing semantic context of their acquisition. BK can help close that gap, as shown in [5]. In the second place, and even more importantly, the ontology matching process is heavy and costly: most of the existing matching tools are complex engineering artefacts comprising a sophisticatedly orchestrated pipeline of matching modules, mapping filtering and semantic verification components. In a recent study [3], we have shown that an appropriately chosen BK source can help significantly lighten the overall matching process. Finally, in specific domains there is a clear need for specific reference knowledge, since the commonly used external knowledge sources, such as WordNet, fail to provide the semantic information that is needed to discover correctly the correspondences between domain specific concepts.

While there is little doubt about the benefits of using BK for ontology matching, outlined as one of the challenges for the field by [1], an important question remains largely unanswered: how to select an optimal BK source for a given ontology matching task out of a set of known BK sources? We understand "optimal" as the source that provides the best quality of the alignment produced for two ontologies. In this paper, we attempt to answer this question by proposing an approach for the automatic selection of a BK source for a given ontology matching task. We situate the problem in an information retrieval framework. The set of known BKs is indexed by using the well-known vector space model while the two ontologies to be aligned are represented as a query document. The comparison between the ontologies and the BKs is based on their content, but also on their structure. We elaborate on the different choices of a similarity measure for this task. Particularly, we show that the commonly used cosine similarity is not the best choice in this scenario and we propose the use of correlation-based similarity measures. The selection system that implements this approach has the properties of being *fully automatic*, *domain independent* and *multilingual*, as well as being entirely *dissociated from the alignment process*.

In order to evaluate the proposed approach, we carry out experiments on benchmark data coming from the ontology alignment evaluation initiative (OAEI). We used as background knowledge sources a mixed set of domain specific and general purpose knowledge graphs. The results show that our approach guarantees the selection of the optimal BK source with respect to each of the matching tasks that has been performed in terms of the quality of the produced alignment by using the selected BK.

The paper is structured as follows. The following section introduces the BK selection problem by focusing on requirements and criteria for selection. Section 3 describes our approach that we support experimentally in Section 4. Several related results are discussed and compared to our method in Section 5 before we conclude in Section 6.

2 The Background Knowledge Selection Problem

Ontology matching is the process of automatically discovering semantic correspondences between the entities of two ontologies that are assumed to cover overlapping domains of knowledge. The result of the ontology matching process is a set of pairs of cross-ontology entities (names of concepts or properties) called an *alignment*, where the entities of each pair are bound by a given semantic relation (most commonly equivalence) [6].

As pointed out in the introduction, using an appropriate background knowledge source (hereafter, BK for short) in the matching process can help improve the results by potentially decreasing the complexity of the process. A BK is understood as any piece of external information that can be used in order to improve the matching quality. According to [1], one can consider as BK for ontology matching a large range of external sources, such as linked data, domain specific corpora of schema and alignments, domain specific or general purpose ontologies, dictionaries and thesauri, lexical databases. Since, on the one hand, the choice of BK has a direct impact on the results and, on the other hand, – there is a multitude of available BK sources, researchers in the field have recognized the need for an approach to select automatically the optimal BK for a given matching task.

2.1 Criterium of Optimality of a BK

There is a large set of BK sources that can be considered, some of them not even known to the user. For that reason, it is important to frame formally the criterium that defines an optimal BK. Since the aim of using a BK is to provide good quality matching results no matter the provenance and nature of the BK, its choice has to be motivated by the maximization of the alignment quality. Suppose that for a given matching task (a pair of ontologies to be aligned). The best BK for this matching task, that we call *optimal* BK, is selected as the one that produces the best alignment.

Let $\Gamma = \{BK_1, ..., BK_n\}$ be a set of known BKs. Let s, t be two ontologies (s for *source* and t for *target*) and $\mathcal{A} = \{A_1, ..., A_n\}$ – a set of alignments of s and t each using a different BK from Γ, given as $A_i = (s, t, m_i, BK_i)$, where A_i is produced by using BK_i, m_i is the metric selection value between (s, t) and BK_i , $\forall i = 1, ..., n$.

Definition 1 (Optimal BK). *We define the* **optimal** *background knowledge source for the task of matching s and t as the background knowledge source $BK_o \in \Gamma$, which corresponds to an alignment from $\mathcal{A}$ with the higher quality. In case multiple BKs produce alignments with maximal m_i values, the one with the lowest number of entities is defined as optimal.*

The optimality criterium given above is inherently semantic, although this remains implicit in the definition. The BK-based ontology matching systems are designed in such manner that the BK that maximizes the quality an alignment of two ontologies is the one that is closest in content to these ontologies. This is the reason why our selection method is based on content similarity between the BK sources and the input ontologies. In our experiments, we show that a BK selected by the help of this method is optimal in the sense of definition 1 (see Section 4).

As a final remark, note that we do not base the selection criterium on the improvement provided by a BK as compared to a direct matching. Our assumption is that an user is looking for a BK because she is not satisfied with the results achieved by a direct matching.

2.2 Requirements to an Automatic BK Selection System

We outline the requirements that, in our view, a BK selection system has to meet.

1. BK **type independence**. The system has to be able to take into account BKs serialized differently (SKOS, OWL, ttl, *etc.*) as long as there exists a parser able to extract the textual information and structure from the BKs and of different semantic nature (thesauri, ontologies, lexical databases, corpora).
2. **Domain independence.** The system should be able to propose a BK for a pair of ontologies of any given domain of knowledge.
3. **Multilingualism.** The system has to be able to select a BK that assists the alignment of cross-lingual ontologies.
4. **Optimality.** It should be guaranteed that the system returns an optimal BK with respect to a matching task in the sense of definition 1.

3 An Information Retrieval Approach to Automatic Selection of BK

We situate the BK selection problem in an information retrieval framework. We consider Γ as a corpus of documents and a given pair of ontologies s and t

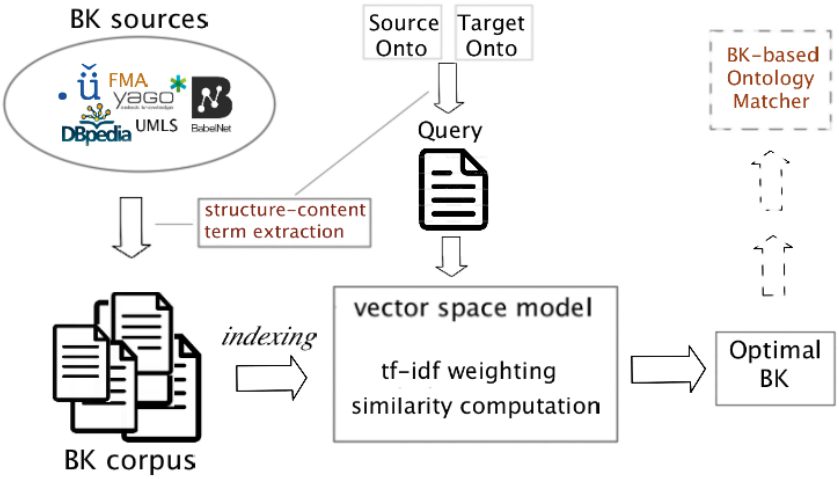

Fig. 1: The BK indexing and selection process. The elements in dashed lines are not part of the selection workflow.

to be aligned – as a query document in the form of a set of terms. The corpus Γ is indexed in order to represent its content by using standard information retrieval techniques, that we explain further on. One of the particularities of our approach consists in the fact that we construct a common index for all known BKs, independent on their domain and focus. As we shall see, in this way we reply to the first two requirements given in the previous section. Additionally, the effort of indexing large ontologies is performed only once, which contributes to the efficiency of the approach. Note also, that the query is given in the form of a unique document representing the pair of ontologies (and not one document per ontology), because we want to retrieve the background knowledge that is common for the two and therefore allows for their reconciliation. In the following, we first explain on how we transform the BK sources and the query ontologies to documents. Then we describe the indexing process and present a set of similarity measures that we use in our approach for retrieving the optimal BK for a given matching task. The overall process is depicted in Figure 1.

3.1 Modeling Ontologies and BKs as Structure-Content Documents

We map a BK source to a text document that we call a BK-document in a manner that allows for taking into account both the content and the structure of the knowledge sources into account. The term extraction method consists in the following. Each token of the labels of the concepts of a given BK becomes a term in the corresponding BK-document. In order to preserve the information relevant to the BK structure, we add a given term to the BK-document every time when it appears in a label of a sub-class or a superclass within the BK

hierarchy or, when it appears in the label of a concept in the domain or the range of a property or, more generally speaking, when it appears in the label of a concept of any relation (synonymy, subsumption, etc).

To illustrate this idea, take a part of a BK where the classes {*Author, Administrator, PhD Student, Professor*} are all in a `subclassOf` relation with the class *Person*. Without the use of the structure, the resulting BK-document would be d_{BK} = {*Person, Author, Administrator, PhdStudent, Professor*} with all the terms having the same weight with respect to their frequency of occurrence. However, the term *Person* seems to be more important than the other terms in the BK as it is a label of their common superclass. Not considering this semantic information affects considerably the computation of weights in the indexing phase. By using the structural information, the BK-document becomes d'_{BK}={*Person, Person, Person, Person, Person, Author, Administrator, PhDStudent, Professor*}.

We model in this manner all known BKs in Γ, resulting in the corpus $D_\Gamma = \{d_{BK_1}, d_{BK_2}, ..., d_{BK_n}\}$. We proceed in the same manner in order to represent a pair of ontologies, t and s, as a query document, denoted by $q_{t,s}$ by creating a document for each ontology as shown above and then merging the two documents into a single one.

3.2 Indexation

Prior to indexing, we apply a standard set of text preprocessing methods, such as normalization of characters and spaces, removing diacritics or accents, deleting numbers, punctuations and stop words, tokenization and lemmatization.

We index the documents in D_Γ by using the well-known vector model. We build an indexation matrix M, which has BK-document vectors as rows and index term vectors as columns. Standardly, the index terms are the terms collected from the BK-documents without repetition after preprocessing. We denote the set of index terms by C_t. Each element w_{ij} of M corresponds to the weight of the term t_j with respect to d_{BK_i}.

In order to compute the weights w_{ij}, we use the well-known TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme that captures the importance of a term both within a single document and in a collection of documents. For a term t_j and a BK-document d_{BK_i}, the weight w_{ij} is calculated as follows:

$$w_{ij} = tf(t_j, d_{BK_i}) * log(\frac{n}{df(t_j)}), \quad (1)$$

where $tf(t_j, d_{BK_i})$ is the frequency of occurrence of the term t_j in the document d_{BK_i} and $df(t_j)$ is the number of documents containing t_j and n is the number of documents in D_Γ.

3.3 Retrieving the Optimal BK

We need a similarity measure of some kind, which allows the system to return to the user the BK with highest similarity to the input ontologies query.

In information retrieval, the most commonly applied similarity measure is the cosine similarity, denoted by *cosine*, given as the normalized dot product of two document vectors. We noticed that in certain cases parameter free measures of (rank) correlation can be applied more successfully than the cosine similarity measure for our particular problem. A measure of correlation expresses the degree of dependence of two random variables. It takes values between -1 and 1, assuming strong correlations for high positive values, strong anti-correlations for high negative values and no dependence for values close to 0. We introduce two of the most popular choices for correlation measures and show how they can serve as similarity measures for two documents [7]. Note that the correlation measures to be presented, just as the cosine measure, are based on the dot product of vectors. In that sense, the cosine similarity can be seen as a correlation, just as the correlations can be seen as similarity measures.

Pearson Correlation. The Pearson's coefficient r between two variables X and Y is calculated from their covariance $cov_{X,Y}$ and their standard deviations σ_X and σ_Y in the following way:

$$r_{X,Y} = \frac{cov_{X,Y}}{\sigma_X * \sigma_Y} = \frac{\sum_i (x_i - m_X)(y_i - m_Y)}{\sqrt{\sum_i (x_i - m_X)^2} \sqrt{\sum_i (y_i - m_Y)^2}}, \tag{2}$$

where m_X and m_Y are the means of X and Y.

In our case, each of the variables X and Y corresponds to either a BK-document or a query document that all live in the same space and are therefore representable by the same type and number of features (i in (2) takes values from 1 to the number of index features). The sets of values $\{x_i\}$ and $\{y_i\}$ correspond to the values of the $tf * idf$ vectors in our vector model. Pearson is used to test the linear dependency of variables.

Spearman Correlation. Spearman's coefficient provides a measure of the correlation of two variables X and Y represented as lists of statistical rankings. In contrast to Pearson, it is applied when a general monotonic relationship is expected between the variables. Formally, it is given as

$$\rho_s = 1 - \frac{6 \sum_i d_i^2}{n^3 - n}, \tag{3}$$

where d_i is the difference between the rank of the ith observation of the variable X and the rank of ith observation of the variable Y, $d_i = rank(x_i) - rank(y_i)$. We transform the $tf * idf$ values to ranks by using the partial order on real numbers. In case of equal values, we assign equal ranks.

BK Selection. For a given corpus of indexed BK-documents D_Γ, a query $q_{s,t}$ and a similarity measure $\sigma \in \{cosine, r, \rho_s\}$, the selected BK is the one whose BK-document maximizes the similarity between the query and the documents in D_Γ, given as

$$D_{BK_{s,t}} = \arg\max_{D_{BK} \in D_\Gamma} \sigma(D_{BK}, q_{s,t}). \quad (4)$$

In case more than one BKs provide a maximal similarity to the query, the one with lowest number of entities is selected.

In case where an ontology matching system uses more than one BK at a time, we can recommend the top-K BKs, with respect to their similarity to the input ontologies. These sources can be further combined – a problem that is not addressed in this paper.

Where no optimal BK is found in the set of BKs, we can easily apply threshold limits to avoid this scenario. In our study, we have considered that always an optimal BK exist in the set of BKs.

We show experimentally in Section 4 that the selected background knowledge source $BK_{s,t}$ corresponding to the BK-document $D_{BK_{s,t}}$ is optimal in the sense of definition 1.

4 Evaluation and Results

In order to evaluate our approach, we have conducted experiments on data coming from the ontology alignment evaluation initiative (OAEI)[2] of year 2015. Our aim is to test if our approach selects the optimal BK among a set of BK sources in two scenarios: selecting a domain specific BK or selecting a general purpose BK. We take a set of BKs Γ = {Yago, DBpedia, BabelNet, DBnary, FMA, Doid, Uberon}, described below. The set Γ includes general purpose sources (Yago, DBpedia, BabelNet and DBnary), as well as several domain specific anatomy and biomedical BKs (FMA, Doid, Uberon).

As described in Section 3, the selected BK for each task has to conform to the optimality criterium given in Definition 1. In order to verify this criterium, we need a F-measure score produced as a result of the comparison of an alignment given by an ontology matching system that uses BK in its process and a reference alignment. For the totality of our experiments, we have used the system LYAM++[3], which is entirely based on BK and has also shown to perform well on the OAEI MultiFarm track on which it participated last year [8]. LYAM++ don't use any complex matching methods regard to the BK. For instance, LYAM++ use $HasSynonyms$ relation presented in given BK to match between two concepts.

Note again that the aim of these experiments is not to show the quality of the ontology matching tools, but to evaluate the performance of the BK-selection approach proposed in this paper. The pairs of ontologies to be aligned, as well as the BKs in Γ are modeled as documents and indexed as described in the previous section. In the retrieval phase, we have tested the performance of the three similarity measures given in Section 3 - the commonly used cosine similarity and the two parameter-free correlation coefficients (Pearson and Spearman).

[2] http://oaei.ontologymatching.org/

4.1 The OAEI tracks

The **Anatomy track** aims at discovering alignments between a human anatomy ontology, part of the NCI Thesaurus[3] and a mouse anatomy ontology. This track is considered as a large-scale matching task because the input ontologies are of a large size and very rich semantically. The **Large Biomedical track** aims at aligning three large bio-medical ontologies, namely FMA, SNOMED and the NCI Thesaurus. Since FMA appears in a query couple, we did not consider this ontology as background knowledge source for this track. The Anatomy and the Large biomedical tracks have been selected to make sure that at each experiment our approach selects from Γ the right BK for the right domain, respectively anatomy and biomedicine.

The **Conference track** contains a dataset of about 15 ontologies from the scientific publication field, together with reference alignments. We have used with the Conference dataset to test if the approach selects the optimal general purpose BK among the mixed set of BKs Γ. Note that, although the Conference data are specific to the scientific publishing domain, they can be considered as general purpose due to the type of concepts that are used to describe this domain, often dealing with common sense knowledge.

Finally, the **MultiFarm track** is derived from the Conference track data by translating the conference ontologies into several different languages with the aim to challenge the performance of cross-lingual ontology matching tools. In our scenario, this track is appropriate for testing whether the selection procedure is able of choosing an optimal multilingual knowledge source for aligning cross-lingual ontologies.

4.2 BK sources

We give a quick overview of the BK sources used in this evaluation. BabelNet [9] is a multi-lingual semantic network and an ontology that has been built by merging different encyclopedic and linguistic resources. The integration of these resources has been conducted automatically. BabelNet appears to be an appropriate choice of a BK for the *MultiFarm* track. DBnary[4] [10] is a multilingual lexical database extracted from Wiktionary, which is also a potentially good candidate for the MultiFarm track. DBpedia [11] is a large multilingual knowledge graph extracted from Wikipedia, covering a multitude of areas such as music, films, people, places, *etc.*. YAGO [12] (Yet Another Great Ontology) is a another large multilingual general purpose knowledge graph extracted from Wikipedia, GeoNames[5] and WordNet. Both YAGO and DBpedia can be considered as candidates for a large variety of general purpose ontology matching tasks. Doid[6] is an open source ontology for the integration of biomedical data associated with human diseases. In our evaluation setting, Doid can be potentially

[3] https://ncit.nci.nih.gov/ncitbrowser/
[4] http://kaiko.getalp.org/about-dbnary/
[5] http://www.geonames.org/
[6] http://do-wiki.nubic.northwestern.edu/do-wiki/index.php

used for aligning Biomedical or Anatomy-related ontologies. FMA (Foundational Model of Anatomy) [13] is the reference ontology for the anatomy field. Similarity, Uberon [14] is a multi-species anatomy ontology. Both ontologies can be used for the Anatomy ontology matching task.

Table 1: Anatomy and Large Biomedical Tasks

	Anatomy				Large Biomedical			
	Cosin	**Pearson**	**Spearman**	**F-m**	**Cosin**	**Pearson**	**Spearman**	**F-m**
BabelNet	0.03	0.06	-0.76	0.05	0.0	0.0	-0.66	0.0
DBnary	0.02	0.07	-0.72	0.0	0.02	0.02	-0.51	0.0
DBpedia	0.01	0.01	-0.70	0.0	0.05	0.06	-0.60	0.0
Doid	0.10	0.15	-0.14	0.66	0.15	**0.14**	0.40	0.53
Uberon	0.26	**0.33**	**0.40**	**0.79**	**0.16**	**0.14**	**0.5**	**0.60**
YaGo	0.01	0.01	-0.20	0.0	0.03	0.03	-0.22	0.0
FMA	**0.30**	**0.33**	0.20	0.46	-	-	-	-

Table 2: Conference and MultiFarm Tasks

	Conference				MultiFarm			
	Cosine	**Pearson**	**Spearman**	**F-m**	**Cosine**	**Pearson**	**Spearman**	**F-m**
BabelNet	0.28	-0.08	**0.73**	**0.61**	**0.30**	**0.39**	**0.44**	**0.49**
DBnary	0.06	-0.01	-0.34	0.46	0.16	-0.20	0.22	0.29
DBpedia	0.1	-0.02	0.12	0.12	0.01	0.03	0.09	0.10
Doid	0.10	0.01	-0.05	0.0	0.0	0.0	0.0	0.0
FMA	0.08	-0.05	-0.07	0.0	0.0	0.0	0.0	0.0
Uberon	0.11	0.0	0.05	0.0	0.0	0.0	0.0	0.0
YaGo	**0.30**	**0.11**	0.31	0.10	0.10	0.09	-0.09	0.11

4.3 Results Presentation.

The results from this series of experiments are presented in Tables 1 and 2. For each BK, we provide the average similarity scores obtained by each of the three similarity measures described in Section 3 and the corresponding average F-measure values obtained in the BK-based alignment. The average values are computed over all pairs of ontologies in each track. The best score achieved by each similarity measure, as well as the highest F-measure value are highlighted. The two correlation coefficients take values in the $[-1, 1]$, which explains the negative numbers. We are interested in strong correlations, corresponding to high similarity. Recall that according to our criterium, the BK selected by a similarity measure is optimal if it guarantees the highest F-measure.

4.4 Results Analysis

As it can be seen from the results in Tables 1 and 2, our approach systematically selects the optimal BK, independently on the track, by using the Spearman

correlation and very conclusively so with similarity scores around and above 0.5. The performance of the selection method is flawed when using the cosine similarity, which is the common choice for a similarity measure in an information retrieval setting, but ranks only second best in our experiments. Namely, it fails to detect the optimal BK on the Anatomy and the Conference tracks and on the other tracks its outcome does not always help to come up with a clear cut decision.

We explain this observation by the fact that the Spearman measure is based on ranks instead of real values and on monotonic dependencies between the two vectors. Precisely, a small variation in the corresponding values of two vectors influences the cosine similarity negatively, while this is less so in case of Spearman. Spearman seems to be also better suited to dealing with highly sparse data (it is most of the times the case that the query document corresponding to two input ontologies contains much less terms than any of the BK-documents in the selection pool).

As for the Pearson correlation coefficient, Spearman appears to largely outperform that, as well. This is mainly due to the fact that the Pearson correlation measure is suited for testing linearity of the variables, while Spearman assumes monotonic behavior of the rank pairs. We draw the reader's attention to the fact on the Anatomy track Pearson yields a maximal value for two different BKs – Uberon and FMA. The BK selected by this measure is Uberon, the smaller of the two sources, conforming to our selection condition given at the end of Section 3. UBERON is well known as the perfect BK for anatomy task. However, if UBERON does not exist in the set of BKs the approach will select FMA as the optimal BK because FMA has the higher correlation value than the other BKs.

Finally, we note that the choice of a similarity measure is more important in the presence of multiple similar in terms of domains BKs, as this is the case in the OAEI experiments, where the only similarity measure that remains unflawed is Spearman.

5 Related Work

We provide an overview of related approaches to ontology matching with regard to two aspects of this task: (1) the use of background knowledge and (2) the automatic selection of background knowledge.

5.1 Ontology Matching Using Background Knowledge

The idea of using background knowledge (BK) for enhancing ontology matching task is not new and has been successfully adopted in several matching approaches. Although not directly relevant to our study, which focuses on the BK selection process, we summarize here the main groupes of relevant approaches.

An intuitive idea is to rely on reusing existing mappings in order to improve the mappings produced by a system. Several approaches [16,17,18] follow this

paradigm. The main drawback of this group of methods is the fact that they depend heavily on the quality of the re-used mappings and, hence, on the performance of the ontology matching techniques that have been used to produce them.

Another approach consists in using a corpus, which can be seen as a rich collection of data elements and their data types, relationships between elements, sample data instances and other information that can be used to discover mappings between entities [19,20]. Furthermore, domain specific ontologies are often seen as quality sources of background knowledge. In [21,22], the alignment process takes place in two steps: *anchoring* and *driving relations*. Anchoring consists in matching the concepts of the source and target ontologies to the concepts of the reference knowledge using standard ontology matching techniques. Relations between source and target concepts are derived by checking if their corresponding anchored concepts are related.

A group of approaches relies on the web in order to discover (by crawling) automatically relations between the input ontologies entities that may exist in various knowledge sources distributed on the web [4].This is particularly useful when the needed background knowledge is spread among different sources. More recently, several approaches have been proposed that rely on general purpose knowledge graphs, such as Yago and DPBedia. It has been shown that such sources are particularly useful for aligning cross-lingual ontologies [5,3].

5.2 Automatic Selection of Background Knowledge

To the best of our knowledge, there are only a few approaches that have addressed the question of automatic BK selection.

In [23], an automatic background knowledge selection approach has been proposed for the particular task of matching biomedical ontologies based on the notion of mappings gain (MG). MG is used to estimate the individual usefulness of background knowledge sources and is defined as a function of the improvement of the number of correct mappings by using a given BK source as compared to a direct mapping of two ontologies. The BK providing the highest MG is selected. The authors have shown experimentally the correlation between the mapping gain and the F-measure in most of the matching tasks that they perform. However, it seems that the notion of MG is not always reliable. Indeed, as reported by the authors, despite having a relatively high mapping gain, using a selected BK may have a negative impact on the results for some tasks; this is the case of using WordNet as BK. The main drawback of the approach is the fact that the selection procedure, defined by the concept of MG, depends on an already existing alignment between the source ontologies.

The closest to our work is reported in [24]. In this approach, a local repository is built from a set of ontologies, to be used as background knowledge sources. These resources are indexed by extracting concept names, comments and labels. In addition and separately, structural features are taken into account. Querying the repository is performed for a given source and target ontology, modeled as sets of key words. If a suitable BK is not found for the given ontologies in the

repository, the method searches the web for appropriate BK ontologies. In the likely case of returning more than one ontology, all found ontologies are used for the matching independently and a unified result set is produced. As reported, the adopted strategy aims at selecting the BK that maximizes the F-measure score for a given matching task, although no experimental results are presented to illustrate and support this selection criterium.

Positioning. We present the major points, which differentiate our technique as compared to the two approaches described above.

Although also relying on information retrieval techniques, in our method, the query is represented as a single document, built in the same way as the BK-documents. This allows to avoid the complex weighted similarity computation in [24], depending on the setting of two parameters. Instead of creating two classes of features (one for terms and one for structure), we embed the structure of the BKs *and* of the input ontologies in their respective textual representations (see Section 3), which has the potential to produce a more compact index.

We make clear distinction between query ontologies and BKs in the evaluation phase. Indeed, the evaluation presented in [24] is highly biased by the fact that the authors use the same type of ontologies as queries and as BK (for example, if aligning two ontologies from the Benchmark track of OAEI, the authors would include the rest of the Benchmark ontologies as BK candidates), which will lead to having in the repository always a very useful BK source at hand. This is however, hardly a realistic scenario. In contrast, we use as BKs well-established and widely used knowledge graphs that are much likely to be called upon as BK sources for solving a real-life alignment problem. In that line of thought, our approach is generic and can handle different domains, contrarily to [23].

In both [24] and [23] the BK selection appears to be strongly coupled with the actual matching procedure. One of the main motivations of work is to dissociate completely the BK selection process from the alignment procedure.

In contrast to our work, the results presented in [24] are not reproducible, because no information is given regarding the selected BKs, neither about the similarity measure that has been used in the process. In turn, no direct experimental comparison to the approach is possible.

We could not make comparison between the works cited in this section by the fact that: - It is difficult to reproduce the scenarios presented in this papers - We do not use the same metrics - Our scenarios are not applicable on these approaches because they do not address the multi-domain selection and the multilingualism problems.

On another hand, the index built in our approach is reusable, no need to construct the index at each ontology matching tasks.

6 Conclusion and Future Work

In this work, we have addressed the problem of automatic selection of background knowledge sources for the task of ontology matching. We propose an

approach using information retrieval techniques implemented in an automatic domain independent system that can handle multilingual input ontologies. We build an index for a set of known, well-established in the semantic web field knowledge sources, that are often used as BK for aligning ontologies. For a pair of ontologies to be matched, the selection process is a result of querying the indexed corpus for semantically similar BK sources from the indexed data. We provide an in-depth empirical study and we show that in certain cases the standard choice of a cosine similarity is not the most optimal and parameter free correlation measures can help discriminate better between close in terms of domains BK sources. We define an optimality criterium of selection based on the quality of the matching and we show experimentally that our approach satisfies this criterium. Contrarily to state of the art approaches, our technique has the advantage of not being based on a preliminary direct matching between the input ontologies.

In the future, we plan to work on optimizing the selection process by improving the quality of the index features. In that respect, we will consider the task of BK preselection for a particular domain, to be applied prior to the selection algorithm. We also plan to improve our selection criterium in order to take into account the trade-off between optimality and efficiency. Finally, we will investigate the benefits of the development of a BK selection method to assist the instance matching and link discovery processes.

References

1. P. Shvaiko and J. Euzenat, "Ontology matching: state of the art and future challenges," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 1, pp. 158–176, 2013.
2. D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, and F. M. Couto, "The agreementmakerlight ontology matching system," in *ODBASE*, pp. 527–541, 2013.
3. A. N. Tigrine, Z. Bellahsene, and K. Todorov, "Light-weight cross-lingual ontology matching with LYAM++," in *ODBASE*, pp. 527–544, 2015.
4. M. Sabou, M. d'Aquin, and E. Motta, "Exploring the semantic web as background knowledge for ontology matching," *J. Data Semantics*, vol. 11, pp. 156–190, 2008.
5. K. Todorov, C. Hudelot, and P. Geibel, "Fuzzy and cross-lingual ontology matching mediated by background knowledge," in *Uncertainty Reasoning for the Semantic Web III*, pp. 142–162, Springer, 2014.
6. D. Ngo, Z. Bellahsene, and K. Todorov, "Opening the black box of ontology matching," in *The Semantic Web: Semantics and Big Data*, pp. 16–30, Springer, 2013.
7. W. J. Conover, *Practical Nonparametric Statistics*. Wiley, 1998.
8. A. N. Tigrine, Z. Bellahsene, and K. Todorov, "Lyam++ results for oaei 2015," *OM at ISWC*, 2015.
9. R. Navigli and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artif. Intell.*, vol. 193, pp. 217–250, 2012.
10. G. Sérasset, "Dbnary: Wiktionary as a lemon-based multilingual lexical resource in RDF," *Semantic Web*, vol. 6, no. 4, pp. 355–361, 2015.

11. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *ISWC 2007 + ASWC*, pp. 722–735, 2007.
12. F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *WWW*, pp. 697–706, 2007.
13. C. Rosse and J. Mejino, "A reference ontology for biomedical informatics: the foundational model of anatomy," *Journal of Biomedical Informatics*, vol. 36, no. 6, pp. 478–500, 2003.
14. M. Haendel, J. P. Balhoff, F. B. Bastian, D. C. Blackburn, J. A. Blake, Y. Bradford, A. Comte, W. M. Dahdul, T. Dececchi, R. E. Druzinsky, T. F. Hayamizu, N. Ibrahim, S. E. Lewis, P. M. Mabee, A. Niknejad, M. Robinson-Rechavi, P. C. Sereno, and C. J. Mungall, "Unification of multi-species vertebrate anatomy ontologies for comparative biology in uberon," *J. Biomedical Semantics*, vol. 5, p. 21, 2014.
15. M. Achichi, R. Bailly, C. Cecconi, M. Destandau, K. Todorov, and R. Troncy, "DOREMUS: doing reusable musical data," in *Proceedings of the ISWC 2015 Posters & Demonstrations Track*, 2015.
16. H. H. Do and E. Rahm, "COMA - A system for flexible combination of schema matching approaches," in *VLDB*, pp. 610–621, 2002.
17. A. Groß, M. Hartung, T. Kirsten, and E. Rahm, "GOMMA results for OAEI 2012," in *Workshop on Ontology Matching*, 2012.
18. B. Saha, I. Stanoi, and K. L. Clarkson, "Schema covering: a step towards enabling reuse in information integration," in *ICDE*, pp. 285–296, 2010.
19. J. Madhavan, P. A. Bernstein, K. Chen, A. Y. Halevy, and P. Shenoy, "Corpus-based schema matching," in *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*, pp. 59–63, 2003.
20. J. Madhavan, P. A. Bernstein, A. Doan, and A. Y. Halevy, "Corpus-based schema matching," in *Proceedings of the 21st International Conference on Data Engineering, ICDE*, pp. 57–68, 2005.
21. Z. Aleksovski, M. C. A. Klein, W. ten Kate, and F. van Harmelen, "Matching unstructured vocabularies using a background ontology," in *EKAW*, pp. 182–197, 2006.
22. Z. Aleksovski, W. ten Kate, and F. van Harmelen, "Exploiting the structure of background knowledge used in ontology matching," in *Proceedings of the 1st International Workshop on Ontology Matching (OM-2006), (ISWC-2006)*, 2006.
23. D. Faria, C. Pesquita, E. Santos, I. Cruz, and F. Couto, "Automatic background knowledge selection for matching biomedical ontologies.," in *PLoS ONE 9(11): e111226. doi:10.1371/journal.pone.0111226*, 2014.
24. C. Quix, P. Roy, and D. Kensche, "Automatic selection of background knowledge for ontology matching," in *SWIM*, p. 5, 2011.