Arabic-English Automatic Ontology Mapping Based on Machine Readable Dictionary

Mustafa Abusalah, John Tait, Michael Oakes

The University of Sunderland
School of Computing and Technology
Sunderland SR6 0DD, UK
+44(0)191 515 3410
{Mustafa, John.Tait, Michael.Oakes }@sunderland.ac.uk

Abstract. Ontologies are the backbone of the semantic web and allow software agents to interoperate effectively. An ontology is able to represent and to clarify concepts and inter-concept relationships and can be used as a framework to represent underlying domain concepts expressed in many different languages. One way to do this is by mapping Ontologies in different languages using an inter-lingual index. In this paper we present a new methodology for ontology mapping in different script human languages (Arabic/English). We identify the steps of extracting concepts on both ontologies and automatically mapping them based on Machine Readable Dictionary (MRD) and Word Sense Disambiguation (WSD) tools. The paper also discusses a unique tool that automatically extracts unmapped concepts and uses MRD and WSD to match them and create semantic bridges between the ontologies.

1 Introduction

In the early stages of the World Wide Web (WWW) it was necessary to develop standards to view web content and to create communication channels. It is no longer necessary to create such channels and inter-applications. However, it is still not possible for applications and agents to interoperate with other applications and agents without having a predefined, human created common framework of the meaning of the information being transferred on both sides. Ontologies alleviate this problem by providing a common framework that allows data to be shared and reused across application, enterprise, and community boundaries [3]. A current focus of activity in the Semantic Web is the construction of domain ontologies. An ontology typically consists of a hierarchical description of important concepts in a domain, along with descriptions of the properties of (the instances of) each concept [2]. Ontology identifies and specifies concepts and concept relations using its main components such as Classes, Relations, Axioms, Functions, and instances. These relations are represented in Ontology languages [1]. As concepts are human language independent in domains like ecommerce, concepts can be mapped to their equivalent concepts in the same human language and to the same concepts in different languages.

This paper describes a new methodology for multilingual ontology mapping. It considers different language scripting schemes and also the implementation of word sense disambiguation (WSD) to improve the mapping process. The mapping process was decomposed in our implementation into three steps, map discovery, map representation (which also includes applying WSD) and map execution.

In the next section we discuss ontology development, selection of ontology development tools and an ontology development language. The third section discusses ontology mapping, semantic bridges, goes through related work, discusses the automatic ontology mapping tool development and implementation. Ontology mapping will produce an inter-lingual index that maps two ontologies at the semantic level. The fourth section covers the evaluation methodology and the experiments. The fifth section concludes our work and recommends future work. The purpose of the ontology mapping approach discussed in this paper is to create a multilingual ontology to be used as a translation base in Cross Language Information Retrieval initially in the travel domain.
2 Development of Arabic English Ontology

Ontology development consumes extensive human effort, especially when working in new domains. The development of a bilingual ontology needs more than linguistic specialists and domain experts in field of the ontology. Ontology development requires development tools, and the extraction of ontology contents needs an ontology parser. All these issues details are discussed in this section in the context of building an Arabic/English cross language retrieval systems for the travel domain.

Ontologies are best used in specific domains [23] [24], and are widely used in medical and ecommerce domains. We selected the travel domain as it is important for tourists of many nationalities and is one of the most important fields of ecommerce. The Tourism and Travel market place is one of the biggest fields of ecommerce applications [3]. Furthermore, less effort is needed to define concepts in Arabic and English in this domain than in more complex domains such as medical applications.

The tool selected to build the ontology was Protégé [19] as it allows the developers to create, browse and edit domain ontologies in a frame-based representation. Protégé has plug-ins to enhance ontology development such as the OWL plug-in, which is used to develop the OWL Ontology and the PROMPT plug-in to help in Ontology mapping.

3 Ontology Mapping

A wide range of expressions are used in this area (merging, alignment, integration, etc). Our understanding of Ontology mapping is that, for each concept defined in the ontology we try to find a corresponding concept in the other ontology [8], if a corresponding concept doesn’t exist in the other ontology we define a corresponding concept using machine readable dictionary MRD and WSD, then align it to the other ontology. In other words Ontology mapping is the process of building a bi-lingual index, to define semantic relations between two ontologies at conceptual level. Ontology mapping translates source ontology instances to the matching target ontology instances. The major contribution of this paper is to present a unique tool we have developed to automatically map two Ontologies written in different script languages Arabic and English and automatically define matching concepts and align them to the ontology in case a corresponding concept was not found.

3.1 Semantic Bridging

Semantic Bridging is the establishment of correspondence between entities from the source and target ontologies. This process is done by creating semantic bridge entities reflecting correspondence between the two ontology entities. Semantic Bridges between Arabic and English Ontologies in our system are provided by defining ontology mapping.

3.2 Related Work

Ontology mapping is one of the most important modules of knowledge management, especially for domain knowledge sharing across communities. We surveyed the research done in the area of information integration and schema matching, with a focus on ontology mapping methodologies and tools. In ontology mapping and merging state of the art we categorize ontology mapping and merging tools into the four groups:
1. Tools for merging two ontologies to produce a new ontology such as: iPrompt, Chimaera [9], and OntoMerge [10].
2. Tools for defining a transformation function that transforms ontology into another such as: OntoMorph [11].
3. Tools for defining a mapping between concepts in two ontologies by finding pairs of related concepts such as MAFRA [12], Prompt, GLUE [13], OBSERVER [16], and FCA-Merge [17].
4. Tools for defining mapping rules to relate only relevant parts of the source ontologies [15].
The new approach suggested in this paper lies under the third category; we extract 2 pairs of related concepts and use them to map the two ontologies. The most well-known tools categorized under the third category above are discussed blow.

1. **MAFRA [12]**: MAFRA mapping process aims to automatically detect similarities of entities contained in two different ontologies. One of the most important contributions of MAFRA framework is the definition of a semantic bridge. This is a module that establishes correspondences between entities from the source and target ontology based on similarities found between them. All the information regarding the mapping process is accumulated, and populates the ontology of mapping constructs, the so called Semantic Bridge Ontology (SBO). MAFRA aims to specify mappings between these two using the semantic bridge ontology. The semantic bridges are defined hierarchically and take into account the structure of the ontologies to be mapped. MAFRA is a monolingual mapping. In our implementation, we used semantic bridging and populated the ontology with mapping constructs, and similar to MAFRA we took into consideration the ontology hierarchy.

2. **PROMPT**: Is available as a plug-in for the open-source ontology editor, Protégé [19]. The tools use linguistic similarity matches between concepts for initiating the alignment process, and then use the underlying ontological structures of the Protégé environment (classes, slots, facets) to inform a set of heuristics for identifying further matches between the ontologies. PROMPTDIFF extension of PROMPT for ontology mapping addresses structure-based comparison of ontologies as its comparisons are based on the ontology structure and not their text serialization. The PROMPTDIFF algorithm works on two versions of the same ontology and is based on the empirical evidence that a large fraction of frames remains unchanged and that, if two frames have the same type and have the same or very similar name, one is almost certainly an image of the other. PROMPT is a monolingual and multilingual mapping tool, but it only maps languages that use the same writing script, such as English and Dutch.

3. **GLUE [13]**: Is a system that employs learning techniques to semi-automatically create semantic mappings between ontologies. GLUE employs a set of machine learning techniques exploiting information in concept instances and the taxonomic structure of ontologies. GLUE uses a probabilistic model to combine results of different learners. The learners that GLUE uses currently rely on ontologies having instances and they work much better if many slot values have text in them rather than references to other instances. GLUE also uses heuristic information related to taxonomy ancestors. GLUE is best used in Monolingual Ontology mappings where both ontologies use the same language.

4. **OBSERVER [16]**: System uses description logic to answer queries using multiple ontologies and information on mappings between them. First, users define a set of inter-ontology relations. The system helps the users with this task by looking for synonyms in the source ontologies. Having defined the mappings, users can pose queries in description-logic terms using their own ontology. OBSERVER then employs the mapping information to pose queries to source ontologies. OBSERVER relies heavily on the fact that descriptions in the ontologies and the queries are intentional and therefore works best in a description logic DL setting. Observer’s main objective is to map two ontologies that have heterogenous vocabulary and having both ontologies in the same language.

5. **FCA-Merge [17]**: Is a method for comparing ontologies that have a set of shared instances or a shared set of documents annotated with concepts from source ontologies. Based on this information, FCA-Merge uses mathematical techniques from Formal Concept Analysis [18] to produce a lattice of concepts which relates concepts from the source ontologies. The algorithm suggests equivalence and subclass–superclass relations. An ontology engineer can then analyze the result and use it as guidance for creating a merged ontology. However, the assumption that two ontologies to be merged share a set of instances or have a set of documents where each document is annotated with terms from both sources is a very strong one and in practice such a situation may occur only rarely. As an alternative, authors suggest the use of natural-language–processing techniques to annotate a set of documents with concepts from the two ontologies. In this technique it is suggested that both ontologies came from the same source, and both are monolingual, as FCA-Merge main aim is to merge domain ontologies to construct one ontology consists of both ontologies concepts. The process includes human assistance.

Many other mapping tools can be found in [20]. All tools discussed above map two ontologies at the conceptual level, some tools uses machine translation algorithms, others uses heuristic algorithms, others compare shared instances, and some tools combine different approaches. But the limitation of these tools is that most of them use monolingual ontologies with heterogeneous taxonomies to map them. And all tools expect to extract mapping relations through existing concepts in both mapped ontologies. In our approach
we create a new equivalent concept if it doesn’t exist after extracting from MRD and applying concepts for a word sense disambiguation tool, and then annotate it to the target ontology. We also consider mapping multilingual ontologies with different script languages.

3.3 Ontology Mapping Implementation in OWL

The OWL vocabulary is rich and very flexible in mapping semantic relations. The owl:sameAs element is used to specify a mapping relationship. owl:sameAs specifies that two resources identified by a URI (Uniform Resource Identifier) refer to the same individual. owl:sameAs is also used to define class equality, which was the owl tag base used to map the Arabic/English Ontologies. Mapping cardinality defines the number of concepts mapped on each side of the ontologies. A concept from the ontology represented as an OWL element can participate in zero, one, or several match correspondence from the other Ontology. In our implementation each concept can participate in zero or 1-1 or N-M cardinalities as each concept in the first Ontology is matched to a zero one or more than one concepts in the second Ontology [6]. For example the English word Accommodation corresponds with the Arabic word سكن or فندق. Figure 1 shows a simple mapping task.

![Figure 1. Simple Concept Matching Task](image)

3.4 Automatic Ontology Mapping Framework

An Ontology mapping process includes several tasks; map discovery, map representation, and map execution define below. Our ontology mapping framework is implemented automatically in the tool that we developed based on J2EE technology [7]. In the following steps we discuss our automatic ontology mapping framework implementation, the mapping framework also specified in figure 2.
Fig. 2. Ontology mapping framework implementation

**Map Discovery**: the process of finding the related concepts or attributes of ontologies and the relations between them. The mapping discovery was done by extracting the English ontology concepts and identifying the unmapped concepts that exists in all OWL nodes at all levels. The concepts exists in owl:class tag, and could be found as a defined concept which is defined as <owl:Class rdf:ID="Diving"> or a reference to a concept defined as <owl:Class rdf:about="#ص afii62829/afii62805"> In both cases the concept is extracted and identified as unmapped if it wasn’t already mapped.

**Map Representation**: identified by equivalence relations between ontologies in a formal way. Semantic Bridges are used to define the mappings between two ontologies. The semantic bridges are described and presented in a meta-ontology. In our implementation we match the extracted unmapped concepts to its relevant translation in the MRD. If the concept was matched in the MRD, the occurring match’s translation are identified and then applied to a WSD. WSD system consists of a list of parallel Arabic English concepts extracted from 50 parallel Arabic English corpus documents of each language. If the concept was match in the disambiguation list, it will be checked for the best translation if not the first occurring match in the dictionary will be selected, if more than one sense is found in the disambiguation list then all senses will be identified for mapping. If the concept was not matched and translation was not found in the MRD, then the concept will stay unmapped. After finishing this process the semantic bridges become ready to be mapped between Arabic and English Ontologies. Map representation process is illustrated in figure 3.
**Map Execution**: In the map execution process concepts defined in the semantic bridges are transformed from the source ontology to the target ontology by evaluating the semantic bridges. If we consider having an unmapped English concept that already has the map representation, then mapping execution is implemented by nesting a child OWL mapping element `owl:sameAs` within the unmapped English concept element. The mapping element also nests a child `owl:Class` element that represent the Arabic equivalent concept. See the English Ontology OWL code portion below:

```
<owl:Class rdf:ID="Diving">
    <owl:sameAs>
        <owl:Class rdf:ID="
 <rdfs:subClassOf>
        <owl:Class rdf:ID="Sports"/>
    </rdfs:subClassOf>
</owl:Class>
```

This process is done on the English Ontology side. On the Arabic Ontology side we check iteratively if the mapped Arabic concept already exists in the Arabic Ontology, if the concept exists then we start the Ontology mapping framework again for a new unmapped concept, if the Arabic concept was not define in the Arabic Ontology then we create a new `owl:Class` element to define the new Arabic concept. Then a new Arabic concept is aligned at the same level the English concept is aligned. This is described in the Arabic Ontology owl code portion below:

```
<owl:Class rdf:about="#
 <rdfs:subClassOf>
        <owl:Class rdf:about="#
    </rdfs:subClassOf>
</owl:Class>
```

As an example of the Ontology mapping framework let us consider the English concept “Diving” was extracted from the English Ontology as unmapped concept, and “Diving” was identified as a sub-class of Sports concept class as it is defined in `rdfs` tag. The Arabic matched concept is aligned to be a sub class of class. The possibility of having mapped concepts that are not aligned at the same level is very rare as this case can only appear if the root concept was not matched in the Dictionary. The mapping process is implemented on both Ontologies, as the above scenario discussed mapping English concept to its Arabic equivalent, the algorithm parses both Arabic and English concepts, when Arabic or English unmapped concept was found. After the mapping process is accomplished, an Inter-lingual index is created on the English Ontology side.

**4 Evaluations**

Ontology mapping can be evaluated in different ways. Our goal is to serve multilingual information retrieval. The ontology mapping evaluation must handle the accuracy of mapping the concepts between the two ontologies [5]. We evaluated the quality of the ontology mapping tool by asking 25 users to annotate 100 concepts to the English ontology then use the tool to map them, then we recorded the users feedback of relevance that indicate whether the mapping of concepts is similar to what they expected. We took into consideration that the match quality may differ from user to user given the same input concepts. We tried to limit this effect to some extent by having feedbacks from different users on the same input concepts, to obtain multiple reliable match results. The concepts source is the English ontology in which it is mapped to Arabic ontology of the same domain. We then asked the users to annotate the English ontology with new
concepts and implement the ontology mapping tool, and used precision and recall measures used in information retrieval systems evaluations to evaluate the quality of mapped concepts. We computed precision and recall as:

\[
\text{precision} = \frac{\text{CorrectFoundMappings}}{\text{AllFoundMappings}}
\]  
(1)

\[
\text{recall} = \frac{\text{CorrectFoundMappings}}{\text{ExistingMappings}}
\]  
(2)

In an ideal state which is common in ontology mappings precision and recall measures, precision=recall=1, in this case correct found mappings = all found mappings = existing mappings, for example if a concept was mapped from English ontology to the Arabic one and there was only a single corresponding concept to be mapped precision=recall=1 and it was correctly mapped.

But neither precision nor recall alone can assess the match quality [6]. In fact precision can be maximized at the expense of poor recall by returning only few correct correspondences, and recall can be maximized at the expense of poor precision by returning all correspondences. Hong-Hi Do et al in [6] suggest using combinational approach Van Rijsbergen’s [25] F-Measure which is also stemmed from information retrieval field. F-Measure is computed as the following equation:

\[
F - \text{Measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]  
(3)

The F-Measure represents the harmonic mean of precision and recall, this evaluation approach was successfully used in [21].

The following chart specifies the three measures results after annotating the English ontology with 100 Concepts and automatically maps them using our proposed tool.

![Measurements comparison chart](chart.png)

Precision average result was 0.83, and Recall was 0.81, while F-Measure average result was 0.82. From the results above it is very clear that the ontology mapped concepts met the user’s expectations. Still we were not able to reach the optimal solution where Precision=Recall=F-Measure=1 as in some cases the system was not able to find a concept match and in other cases the mapped concept was not what the users expected.
5 Conclusions and Future Work

Ontologies have many important uses as they are the backbone of the semantic web. The main purpose of the ontology is to provide a clarification of concepts through concepts relations, and specifications. This leads us to create Ontologies in many languages and map those using concept relations. Much research was carried out in the field of schema alignment and Ontology mapping. Most of the research considered mapping heterogeneous ontologies monolingual some other developed a semi automatic ontology mapping tools multilingually. Other tools developed automatic mappings multilingual but doesn’t meet all our development requirements and didn’t consider different script languages. In our new approach we facilitate mapping of two ontologies to serve Arabic to English CLIR. Concepts were discovered to be mapped and also aligned a new concepts to the second ontology if the equivalent concept doesn’t exist, considering the taxonomy hierarchy. The Automatic ontology mapping tools produced high precision and recall, the reason is that we are evaluating concepts mapping matched in the MRD and refining it by applying WSD. The results showed that using such tool produces high semantic bridges quality. This result is based on mapping 100 concepts only with the increase number of concepts we expect less semantic bridging quality.

Mapping discovery is the process of collecting all unmapped concepts at all levels of the ontology. This was done iteratively as we know the OWL ontology nodes length and levels as its structured. While in the map representation process we were not able to find any previous applications developed having the semantic bridges matches our requirements, most of the literature was about aligning a thesaurus for an Arabic equivalence for the EuroWordNet [22]. Another problem which might not directly related to the translation issue lies in the mapping execution. For instance a 3 star hotel in the UK might be equivalent to a 4 star hotel in Morocco. This problem was not dealt with as the aim of the mapping is to match concepts of both languages, at a semantic level using MRD and WSD as a translation medium. In the present work we will consider Part Of Speech (POS) tagging, text parsers, and Ontology automatic annotations.

Concept extraction and discovery of semantic bridges still depends on many factors that need to be studied, and can lead to new concept mappings. Applying WSD to MRD translation result refines concepts mapping, having a larger parallel corpus and extracting larger set of concepts might improve concepts mapping to obtain a better precision, recall, and F-Measure results.

6 References