Web Semantics: Science, Services and Agents on the World Wide Web 15 (2012) 62–68

Contents lists available at SciVerse ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

MultiFarm: A benchmark for multilingual ontology matching

Christian Meilicke[c,*], Raúl García-Castro[d], Fred Freitas[a], Willem Robert van Hage[b], Elena Montiel-Ponsoda[d], Ryan Ribeiro de Azevedo[a], Heiner Stuckenschmidt[c], Ondřej Šváb-Zamazal[e], Vojtěch Svátek[e], Andrei Tamilin[f], Cássia Trojahn[g], Shenghui Wang[b]

[a] *Universidade Federal de Pernambuco, Brazil*
[b] *Vrije Universiteit Amsterdam, Netherlands*
[c] *University of Mannheim, Germany*
[d] *Universidad Politécnica de Madrid, Spain*
[e] *University of Economics, Prague, Czech Republic*
[f] *Fondazione Bruno Kessler, Trento, Italy*
[g] *INRIA, Grenoble, France*

ARTICLE INFO

Article history:
Received 1 September 2011
Received in revised form
11 April 2012
Accepted 11 April 2012
Available online 27 April 2012

Keywords:
Ontology matching
Benchmarking
Multilingualism
Data integration

ABSTRACT

In this paper we present the MultiFarm dataset, which has been designed as a benchmark for multilingual ontology matching. The MultiFarm dataset is composed of a set of ontologies translated in different languages and the corresponding alignments between these ontologies. It is based on the OntoFarm dataset, which has been used successfully for several years in the Ontology Alignment Evaluation Initiative (OAEI). By translating the ontologies of the OntoFarm dataset into eight different languages – Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish – we created a comprehensive set of realistic test cases. Based on these test cases, it is possible to evaluate and compare the performance of matching approaches with a special focus on multilingualism.

1. Motivation

Ontologies have been introduced in computer science as a means for solving the problem of interoperability between different knowledge sources [1]. In the context of the Semantic Web, it became clear that ontologies do not really solve the problem of semantic interoperability but rather lift it to a higher level of representation. As an answer to this, ontology matching has been established as a field of research concerned with the development of methods for determining equivalent elements in different ontologies [2]. One of the insights of this new field of research is that there is not a single best solution to the problem, but that the performance of a matching method depends on the nature of the ontologies to be matched. Thus, the systematic evaluation of matching methods is an important task. It can reveal strengths and weaknesses of existing methods and guide the selection of the most appropriate method for a given task.

In the past six years, the OAEI has carried out systematic evaluation of ontology matching technology, providing many important insights [3]. While the OAEI features a variety of different benchmark datasets covering a wide range of typical matching problems, almost all datasets considered so far assume that the ontologies to be aligned use English as a common language for naming and describing concepts and relations. This assumption is significant as virtually all matching methods are based on a lexical matching step in which the names of elements are compared, providing an initial estimate of the likelihood that two elements refer to the same real world phenomenon [2].

The increased awareness of the usefulness of ontologies for practical applications has lead to a situation where an increasing number of ontologies actually used in real world applications do not use English as a base language. As argued by Fu et al. [4], such ontologies are an important link between the information available on the Semantic Web and the individual user that prefers to have information presented in his or her local language. The existence of such multilingual ontologies pushes the ontology matching problem to a new level as the basic step used by most matching algorithms has to be completely revised. However, currently there have only been a few attempts to tackle the problem of multilingual ontology matching (e.g. [5–8]).

We think that further progress in this area is hindered by the lack of a commonly accepted benchmark dataset with a

[*] Corresponding author.
E-mail address: christian@informatik.uni-mannheim.de (C. Meilicke).

1570-8268/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.websem.2012.04.001

special focus on multilingualism. This view is supported by the observation that existing publications on the topic always rely on a very specific dataset for evaluation that has been created for the purpose of the publication and that have serious shortcomings which are described in more details below. The existence of a carefully engineered and commonly accepted benchmark dataset would be an important enabler fostering progress in multilingual ontology matching in the same way, as the current OAEI datasets have fuelled research in monolingual ontology matching.

In this paper, we attempt to solve the problems described above by proposing a comprehensive benchmark dataset for multilingual ontology matching. This dataset has been jointly created by the authors on the basis of an existing dataset from the OAEI campaigns. The proposed benchmark consists of seven ontologies for which mutual reference alignments have been created manually. Each of the ontologies has been translated into eight different languages other than English—Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish. Each combination of ontologies and languages establishes a test case for multilingual ontology matching summing up to roughly 1500 test cases.

The rest of the paper is structured as follows. We first describe characteristics to be taken into account when defining a multilingual dataset (Section 2). We discuss existing multilingual datasets and evaluations pointing to problems that limit the validity of these datasets for evaluation purposes (Section 3). We present the MultiFarm dataset providing details about generating translated ontologies as well as creating the reference (gold standard) alignments between the ontologies (Section 4). Then, we focus on some decisions we made while creating the dataset both in terms of language-independent and language-specific aspects (Section 5). In a preliminary series of experiments, we evaluated current state-of-the-art matching systems against the dataset (Section 6). Finally, we comment on the availability of the dataset and conclude with a discussion of remaining shortcomings and future possible improvements (Section 7).

2. Characteristics of a multilingual dataset

In this section, we present different characteristics to be taken into account when defining a multilingual dataset, since they can affect the results of an ontology matcher. Most of the listed features could also influence a monolingual alignment task, as they are mainly related with the Natural Language (NL) descriptions associated to ontology elements, and the ontology structure *per se*. The identified characteristics have been distributed into three levels: (a) Format or encoding level; (b) Lexical and terminological level; and (c) Ontology structure level. Without claiming to be exhaustive, the set of characteristics accounted for in this section covers those aspects of the NL descriptions associated to ontologies as well as ontology expressiveness. They are currently supported by the most commonly used ontology formalisms. We argue that the presence or absence of these ontological features will contribute in the success of the alignment task.

2.1. Format or encoding level

This level includes those characteristics related to the encoding in which the ontology is serialized, the alphabet used in the labels or NL descriptions associated to ontology elements, and the format used for labels.

- *Encoding*. The character encoding in which the ontology is serialized (e.g., UTF-8) can affect the alignment task, as some tools can process multiple encodings, whereas others cannot.
- *Diacritics*. This feature specifies whether diacritics are used in labels or any other type of NL descriptions associated to ontologies. In some languages, the same word written with or without accent can have a different meaning (e.g., in Spanish 'río' means river, whereas 'rio' without accent refers to the first person singular of the verb 'laugh').
- *Language tags*. In specific syntaxes, such as RDF/XML, one can restrict the scope of a particular label (or any other type of NL description related to the ontology) to a certain natural language (e.g., '@en' for English). At a multilingual level, such a language tag may also contribute to avoid errors, since certain groups of languages with common roots share the same words with different meanings (e.g., 'nombre' in Spanish means 'name' and in French 'number').
- *NL description placement*. This characteristic has to do with the place where NL descriptions of ontology elements appear: in URIs, in labels (using *rdfs:label*, *skos:preflabel*, etc.), in both places, or in an external linguistic model created for that purpose (see LIR,[1] LexInfo, lemon[2]). Identifiers in URIs suffer from some restrictions of the URI naming scheme (e.g., some characters such as white spaces cannot be part of URIs).
- *Word separation*. Specifies the way used to separate words in multiple-word terms in URIs or as label annotations (e.g., CamelCase, hyphen, white space). A correct identification of the multiple words that compose a term is necessary to avoid mistakes (e.g., 'hasVAT' consists of the verb 'has' and the acronym 'VAT').
- *Capitalization*. Specifies how capital letters are used in labels or terms (only first word capitalized, all words capitalized, etc.). In some cases, capitalization may lead to incorrect matchings (e.g., 'white house' vs. 'White House').
- *Punctuation*. When showing up in NL descriptions (mostly, compound words or complex Noun Phrase constructions), punctuation marks may signalize the several components that make up a term (e.g. 'Acquisitions through business combinations, intangible assets').

In Section 5 we explain which of these features appears in MultiFarm.

2.2. Lexical and terminological level

This level includes those characteristics concerning the linguistic descriptions that may be related to ontology elements. The amount and type of linguistic descriptions range from labels and comments (as supported by the RDF/XML syntax) or terminological variants (such as the ones enabled by SKOS properties), to more complex linguistic descriptions.

- *Terms as ontology labels*. Specifies whether terms are provided for naming ontology elements. We understand terms as words or expressions that have a precise meaning in a certain domain. When dealing with general knowledge ontologies, we could talk about lexical entries.
- *Definitions*. Specifies whether labels (terms, lexical entries) are accompanied by definitions or glosses in natural language. These definitions can be used in the alignment task to disambiguate the meaning of terms, as they usually provide contextual information (e.g. reference to the superclass, specific properties of the term, etc.).

[1] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/downloads/63-lir.

[2] LexInfo and lemon are available from http://lexinfo.net/.