Multilingual Ontology Mapping: Challenges and a Proposed Framework

(Extended Abstract)

Bo Fu[1], Rob Brennan[1] and Declan O'Sullivan[1]

Abstract. A key problem in supporting multilingual information retrieval and digital content management is reasoning about overlapping context domains. Ontologies are currently emerging as representation techniques for overlapping complimentary context domains. To date, research has focused on the mappings of monolingual ontologies, however, the issue of mapping ontologies written in different natural languages is relatively unexplored at the moment. This paper discusses challenges in the area of multilingual ontology mapping and proposes the semantic oriented mapping for multilingual ontologies (SOMMO) framework to advance the state of the art in multilingual ontology mapping. The SOMMO framework aims to improve multilingual ontology mapping results generated from existing monolingual ontology matching techniques by evaluating the semantics embedded in both the source and target ontologies.

1 INTRODUCTION

In recent years, ontologies have gained a large amount of attention as a part of the process of achieving semantic interoperability. Usage of ontologies traverses many disciplines, in Agriculture, the Agricultural Ontology Service[2] from the Food and Agriculture Organization (FAO) provides reference standardization for defining and structuring Agricultural terminologies. Since all FAO official documents must be made available in Arabic, English, Chinese, French and Spanish, a large amount of research has been carried out in the translations of large multilingual agricultural thesauri [1], mapping methodologies for them [2] [3] and a definition of requirements to improve the interoperability of these multilingual information resources [4]. In education, the Bologna declaration has introduced an ontology-based framework for qualification recognition [5] across the European Union, in an effort to best match labour markets with employment opportunities. In e-learning, educational ontologies are used to enhance learning experience [6], and to empower system platforms with high adaptivity [7]. In finance, ontologies are used to model knowledge in the stock market domain [8] and portfolio management [9]. In medicine, ontologies are employed to improve knowledge sharing and knowledge reuse, for example, a notable amount of research has focused on the creation of a traditional Chinese medicine ontology [10].

Usage of ontologies grew not only in terms of the number of application domains but also in their choices of natural languages as researchers across borders began to build domain specific knowledge bases. Reasoning and mapping of these multilingual ontologies thus has become a pressing issue. Given large and complex multilingual ontologies, it is unlikely that the mapping process would be practical if solely based on human processing, therefore, fully/semi-automated multilingual ontology mapping systems are needed.

The concept of creating ontologies that comprise different natural languages was explored when Carpuat et al [14] merged thesauri that were written in English and Chinese into one bilingual thesaurus in order to minimize repetitive work while building ontologies. A language-independent, corpus-based approach was employed to merge *WordNet*[3] - written in English, and *HowNet*[4] - written in Chinese by aligning synsets from the former and definitions of the latter. Similar research [15] has been done to match Dutch thesauri to *WordNet* by using a bilingual dictionary, and concluded a methodology for vocabulary alignment of thesauri written in different languages. Such methods succeed in aligning large numbers of words, however, they do not take structural aspects into account. Due to the nature of thesauri - being large collections of words, definitions and synonyms - ignoring their structures when generating a mapping poses little problem. Given the more complex structure and sophisticated class relationships of ontologies, such a method would be insufficient as the structures of these ontologies cannot be over-looked to form accurate mapping results.

Espinoza et al. [16] demonstrate a tool – *LabelTranslator* to empower end-users with choices of natural language when gaining knowledge from a given ontology, it is designed to ensure information represented in an ontology using one particular natural language would still achieve the same level of knowledge expressivity if translated into another natural language. The name, *LabelTranslator* is self-explanatory, it translates labels in a given ontology into one of three natural languages, English, Spanish and German. Users are allowed to select any label in a given ontology for *LabelTranslator* to translate, which then returns the selected term's translation along with its namespace and description. The system is comprised of seven steps: users must first tell the system which labels to translate; the system then translates the selected terms using lexical resources and translation web services, if compound

[1] Knowledge and Data Engineering Group, School of Computer Science and Statistics, Trinity College Dublin, Ireland. Email: {bofu, rob.brennan, declan.osullivan}@cs.tcd.ie.

[2] http://www.fao.org/aims/aos.jsp

[3] http://wordnet.princeton.edu

[4] http://www.keenage.com

words are presented, they will be split into components for the translators; for each translated term, the system obtains a list of senses which is then used for disambiguation; in order to return the most appropriate translations, *LabelTranslator* determines the context by retrieving sets of other labels that are associated with the selected terms; it then lists senses for these labels in context; for each candidate translation, their senses are ranked by comparisons made to the context senses to produce a rank list; once the correct sense is selected from this ranked list, it finally updates the linguistic information of the ontology. In *LabelTranslator*, labels are selected one at a time by the user, the translation and description of the selected label are then presented. In the process of translating labels to a preferred natural language other than the original one, it aids the user to better understand the subject area. While *LabelTranslator* highlights challenges when translating multilingual ontologies automatically and includes sophisticated sense disambiguation mechanisms, it is built to translate an ontology from one natural language into another so that it is human readable, however, it does not deliver translated machine readable ontology documents so that software agents could manipulate and annotate.

Pazienza & Stellato [17] propose a linguistically motivated approach to ontology mapping, the framework urges the usage of linguistically enriched expressions when building an ontology and envisions systems that can automatically discover the embedded linguistic evidence and establish alignments that support users to produce sound ontology mapping documents. A three-step methodology is proposed where sets of ontologies with readily embedded linguistic resources are built at the ontology development stage and are fed to the automatic mapping system. A plug-in, *Ontoling* was also developed for the ontology editor *Protégé*[5] that enables users to browse linguistic resources provided by *WordNet* and *FreeDict*[6] during the ontology creation process. Though this methodology promises improved mapping results, the multilingual enriched ontologies demanded by the framework are hard to come by when such specifications are not currently included in the OWL standardization [18] effort.

A large amount of research has been done in the area of monolingual ontology mapping [11], however, the concept of multilingual ontology mapping is relatively new. Matching contests such as the Ontology Alignment Evaluation Initiative[7] (OAEI) encourage the progression of automated ontology matching tools and recognize the importance of addressing issues that are associated with multilingual ontologies. In the most recent OAEI competition, a test scenario involving the mapping of web site directories written in English and Japanese was defined [12]. Among thirteen contestants, four took part in this test scenario, however, only one matching tool was able to submit results [13] to the program.

The main challenges for (semi-)automated multilingual ontology mapping and a proposed framework is discussed in the following section.

2 CHALLENGES IN MULTILINGUAL ONTOLOGY MAPPING & THE SOMMO FRAMEWORK

In a scenario where automated mapping of two ontologies that are written in different natural languages is desired, one approach to achieve such a process is by translating one of them into the natural language that is used by the other ontology, e.g. using machine translation techniques, before applying monolingual ontology matching techniques. In such a multilingual ontology mapping approach, challenges are mainly found in the ontology translation phase and the monolingual ontology matching phase.

Being able to identify the most appropriate translation results of ontology concepts is crucial in the ontology translation phase. It is the author's opinion that these translated concepts will hugely impact on the quality of ontology mapping results generated by existing matching tools, since lexical matching techniques currently tend to dominate in the most successful matching tools [12]. Regardless of recent advances in the development of monolingual ontology matching tools, challenges remain in the generation of accurate matching results. Among ten challenges identified by Shvaiko & Euzenat [19] in the field of ontology matching, the discovery of background knowledge of a specific ontology is an important issue, most recent progress as discussed in [20] [21] attempts to resolve this critical matter.

To address the aforementioned challenges, the semantic oriented mapping for multilingual ontologies (SOMMO) framework is proposed. For each class, instance and property in a source ontology that is to be translated, a collection of translation candidates can be generated using existing machine translation tools such as the *GoogleTranslate* API[8] and the SDL *FreeTranslation* online translator[9]. Using lexicon dictionaries and based on the knowledge represented in the target ontology, a target lexicon database can be created which stores sets of synonyms for all the target concepts. In order to choose the most preferred translation results for each source concept, the target lexicon data-store is used to influence the translation selection algorithm. The source translation candidates are first compared against the sets of synonyms, if matches are found, for each matched target term and/or synonyms, their immediate surrounding terms, i.e. semantics – parent, child, sibling – are collected, and are ranked based on the similarity of their surrounding terms to that of the source terms. The highest ranked target term will be chosen as the most preferred translation result for the source term. If no matches are found when the candidates and synonyms are compared, or when surrounding term comparisons conclude no similarities, the translation selection is solely based on the semantic representations of the source term. In such a case, for each translation candidate, a set of interpretive keywords can be collected which describe the meanings of these candidates using a dictionary. These keywords can then be compared to the surrounding terms of the source term. Based on matches of these keywords, translation candidates can be ranked, with the highest ranked candidate being chosen as the most

[5] http://protege.standord.edu
[6] http://www.freedict.com
[7] http://oaei.ontologymatching.org
[8] http://code.google.com/p/google-api-translate-java
[9] http://www.freetranslation.com
[10] http://jena.sourceforge.net
[11] http://alignapi.gforge.inria.fr

preferred translation result. If no keywords match the source's surrounding terms, a translation result is generated by an automated machine translator. Using tools such as the Jena framework[10], the source structure can be rebuilt, together with the translated entities and expressions of the source concepts, a translated source ontology document can be created.

Given the target ontology and a translated source ontology – now both represented in the same natural language – matching relationships can be determined by applying existing monolingual ontology matching techniques such as the *Alignment* API[11]. Finally, the metadata gathered during the ontology translation phase are of use in the final monolingual matching process, since relationships were already established between some source terms and target terms when the latter are used to influence the translation outcomes of the former. Together with existing monolingual ontology matching tools, such metadata can assist the rendering of more accurate and higher confidence matching results between the translated source ontology and the target ontology. Hence reliable matching relationships are generated between concepts from the original source ontology and the target ontology regardless of their natural languages originally used.

The development of the SOMMO framework is part of on-going research work and several test cases are being designed to evaluate such an approach.

ACKNOWLEDGMENT

This research is partially funded by Science Foundation Ireland (SFI) as part of the National Development Plan (NDP) 2007-2013.

REFERENCES

[1] Chang C. and Lu W., The Translation of Agricultural Multilingual Thesaurus, in Proceedings of the 3rd Asian Conference for Information Technology in Agriculture, 2002.

[2] Liang A., Sini M., Chang C., Li S., Lu W., He C. and Keizer J., The Mapping Schema from Chinese Agricultural Thesaurus to AGROVOC, 6th Agricultural Ontology Service (AOS) Workshop on Ontologies: the more practical issues and experiences, 2005.

[3] Liang A. and Sini M., Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures, New Review of Hypermedia and Multimedia pp. 51 -- 62, 12 (1) 2006.

[4] Caracciolo C., Sini M. and Keizer J., Requirements for the Treatment of Multilinguality in Ontologies within FAO, Food and Agricultural Organisation of the United Nations, 2007.

[5] Vas R., Educational Ontology and Knowledge Testing, The Electronic Journal of Knowledge Management of Volume 5 Issue 1, pp. 123 -- 130, 2007.

[6] Cui G., Chen F., Chen, H. and Li S., OntoEdu: A Case Study of Ontology-based Education Grid System for E-learning, The Global Chinese Conference on Computers in Education conference, 2004.

[7] Sosnovsky S. and Gavrilova T., Development of Educational Ontology for C-programming, International Journal "Information Theories and Applications" Volume 13, pp. 303 – 308, 2006.

[8] Alonso L. S., Bas L. J., Bellido S., Contreras J., Benjamins R. and Gomez M. J., WP10: Case Study eBanking D10.7 Financial Ontology, Data, Information and Process Integration with Semantic Web Services, FP6-507483, 2005.

[9] Zhang Z., Zhang C. and Ong S. S., Building an Ontology for Financial Investment, in Intelligent Data Engineering and Automated Learning - IDEAL 2000, Data Mining, Financial Engineering, and Intelligent Agents, Second International Conference, pp. 308 – 313, 2000.

[10] Fang K., Chang C. and Chi Y., Leveraging Ontology-Based Traditional Chinese Medicine Knowledge System: Using Formal Concept Analysis, in Proceedings of the 9th Joint Conference on Information Sciences, 2006.

[11] Euzenat J. and Shvaiko P., Ontology Matching, Springer 2007.

[12] Multilingual Directory Data Set Specification, http://ri-www.nii.ac.jp/OAEI/2008, last accessed December 2008.

[13] Ontology Alignment Evaluation Initiative 2008 Results, http://oaei.ontologymatching.org/2008/results, last accessed December 2008.

[14] Carpuat M., Ngai G., Fung P. and Church W. K., Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet, In Proceedings of the 1st Global WordNet Conference, 2002.

[15] Malaise V., Isaac A., Gazendam L. and Brugman H., Anchoring Dutch Cultural Heritage Thesauri to WordNet: Two Case Studies, Proceedings of the Workshop on Language Technology for Cultural Heritage Data, 2007.

[16] Espinoza M., Gomez-Perez A. and Mena E., LabelTranslator – A Tool to Automatically Localize an Ontology, ESWC 2008, LNCS 5021, pp. 792 – 796, Springer 2008.

[17] Pazienza T. M. and Stellato A., Linguistically Motivated Ontology Mapping for the Semantic Web, Semantic Web Applications and Perspectives 2005, second Italian Semantic Web Workshop, 2005.

[18] OWL 2 Web Ontology Language Profile, http://www.w3.org/TR/2008/WD-owl2-profiles-20081202, last accessed December 2008.

[19] Shvaiko P. and Euzenat J., Ten Challenges For Ontology Matching, in Proceedings of the 7th International Conference on Ontologies, DataBases and Applications of Semantics (ODBASE) 2008.

[20] M. Sabou, M. d'Aquin and E. Motta, Exploring the Semantic Web as Background Knowledge for Ontology Matching, Journal on Data Semantics XI, LNCS Vol. 5383, pp. 156-190, Springer 2008.

[21] F. Giunchiglia, P. Shvaiko and M. Yatskevich, Semantic Matching, Encyclopedia of Database Systems, 2009, to appear.