Evaluation of a Semantic-Oriented Approach to Cross-Lingual Ontology Mapping

Bo Fu, Rob Brennan, Declan O'Sullivan
Knowledge and Data Engineering Group, School of Computer Science and Statistics,
Trinity College Dublin, Ireland
{bofu, rob.brennan, declan.osullivan}@cs.tcd.ie

ABSTRACT
Most ontology mapping research has focused on the matching of ontologies written in the same natural language, and developing tools and techniques that support this monolingual ontology mapping process. However, as knowledge modelling is not restricted to the usage of a single natural language, mapping systems must be able to operate upon ontologies that are labelled in diverse natural languages. This paper outlines a semantic-oriented cross-lingual ontology mapping framework that makes use of several information sources to influence the selection of ontology label translations in the process of generating high quality mapping results, and presents a high-level overview of the evaluation strategy of the proposed framework.

Keywords
Cross-Lingual Ontology Mapping; Appropriate Ontology Label Translation; Multilingual Ontologies.

1. INTRODUCTION
Benjamins et al. [1] identify multilinguality as one of the great challenges for the semantic web, and point out that one way to address this challenge is by providing assistance for the annotation of ontologies regardless of the natural languages used in them. However, to date, research in the field of ontology mapping has largely focused on the matching of ontologies labelled in the same natural language, where various monolingual ontology matching techniques have been developed as documented by Euzenat & Shvaiko [2]. With ontologies being widely accepted as a knowledge management mechanism in multilingual organisations [3] and used in a range of applications including machine translation [4], information retrieval (IR) [5] and cross-lingual IR [6], multilinguality is increasingly evident in ontologies. One way to enable knowledge discovery, sharing and reuse across natural language barriers in ontology-based systems is by means of cross-lingual ontology mapping (CLOM).

This paper proposes the semantic-oriented cross-lingual ontology mapping (SOCOM) framework and presents a high-lever overview of its evaluation.

EKAW 2010, October 11-15, 2010, Lisbon, Portugal.

2. THE SOCOM FRAMEWORK
The semantic-oriented cross-lingual ontology mapping (SOCOM) framework is designed specifically for cross-lingual mapping tasks carried out in multilingual environments. In doing so, it first transforms one of the given ontologies into an equivalent of itself that is labelled in the natural language used by the other(s), it then applies existing monolingual matching techniques. The transformation of an ontology requires the translation of ontology labels from the source natural language to the target natural language, whereby the notion of appropriate ontology label translation (AOLT) is employed. An AOLT is a translation that is most likely to maximise the success of the subsequent monolingual ontology matching step. The AOLT selection process therefore is concerned with identifying the translations that will most likely enhance the matching ability of monolingual matching techniques, but not necessarily the translations that are linguistically most correct.

To achieve AOLT, several sources of information are used. Firstly, the *source ontology semantics* are used to indicate the context of use for the to-be-translated resource labels. Given a certain position of a node, the labels of its surrounding nodes (i.e. context) can be analysed. For example, for a class node, the labels of its super/sub/sibling-classes can illustrate its context of use. Secondly, since the source ontology is transformed so that it can be best mapped to the target ontology, the *target ontology semantics* can be perceived as translation selection guidelines. For example, when several candidate translations are linguistically correct for a label, its AOLT is the one that is closest to what is used in the target ontology. Thirdly, *mapping intent* captures the user's motive in a CLOM scenario. For example, when working in a highly refined domain such as medicine, achieving highly precise matches would be priority. Whereas when merging knowledge repositories, gaining reasonable recall in the matches generated may be desired. With known intent, the SOCOM framework selects the most suitable translation source(s) in order to generate mappings with high precision and/or recall. Fourthly, *background knowledge* can be drawn on the ontology domains which can be system specified or user specified. In other words, encyclopedia or users can assist the AOLT process by providing additional context of use. Fifthly, to draw on *user expertise*, the SOCOM framework allows a user to specify preferred translation sources and/or matching algorithms. Sixthly, *mapping assessment* is used as a feedback mechanism in the SOCOM framework, whereby statistics containing top-rated translation sources and/or matching techniques are collected to aid the future execution of the framework. This feedback can be implicit or explicit. Implicit feedback is generated when the system assumes certain matches

are correct and identifies the most effective tools based on the assumption. Explicit feedback is generated by the users and is more reliable. Seventhly, *time constraints* may limit the run time for the AOLT process. E.g., when rapid execution is desired, the user can turn on/off certain features dynamically. Lastly, not all of the aforementioned resources will be always available to every CLOM scenario. *Resource constraints* therefore may restrict the level of sophistication of the AOLT selection process.

3. EVALUATION STRATEGY

A state of the art review is conducted first to identify current approaches to CLOM. Through this review process, a generic approach to CLOM was identified and implemented that uses off-the-shelf machine translation tools and monolingual ontology matching techniques. To investigate the effectiveness and to identify potential limitations of this generic approach to CLOM, it is evaluated in two CLOM scenarios involving ontologies written in Chinese, English and French. These ontologies contain approximately one hundred entities and are of the semantic research community and the bibliography domain. Results from these experiments showed that mappings can be neglected by monolingual matching tools when entity labels are translated independently from the ontologies of interest. When the translations of ontology labels are carried out in isolation of the CLOM tasks at hand, inadequate and synonymic translations can introduce further complications to the subsequent monolingual matching step.

Based on this finding, the notion of appropriate ontology label translation arose. An initial framework prototype is implemented that makes use of the readily defined semantics of the given ontologies in a CLOM scenario. This prototype is evaluated against the generic approach in the aforementioned CLOM scenarios using the same multilingual ontologies and gold standards. Experimental results showed that the SOCOM framework generated higher quality mapping results than the generic approach due to its ability to select translations that are similar to what were used by the target ontology in a specific CLOM setting.

Motivated by this initial result, a second framework prototype was then designed and implemented to draw on additional inputs (discussed in section 2) in the AOLT selection process, effectively allowing fine tuning of the system. This second prototype is evaluated against the generic approach in the same CLOM experiments involving the aforementioned multilingual ontologies. Various combinations of the AOLT influence sources were executed in a range of experimental runs of the framework, and several sets of mappings were generated. Versatility in these mapping results demonstrated the flexibility of the AOLT selection mechanism and showcased the tuning ability of the SOCOM framework.

Furthermore, as the experiments discussed above only concern ontologies of relatively small sizes, to assess the scalability of the framework, the second prototype was applied in a real-world CLOM setting involving large organisational ontologies written in English and German. These ontologies contained over 7000 entities and were generated semi-automatically using enterprise data of the technical customer support domain. More details of how these ontologies are generated can be found in [7]. Mappings were then generated using the SOCOM framework between these large multilingual ontologies in English and German. These mapping results then enabled cross-lingual document retrieval of an adaptive personalised result composition and presentation system. Bilingual users can issue queries in German and retrieve relevant as well as personalised content in English. More details of this information retrieval and composition system can be found in [8].

Lastly, in all the experiments carried out, precision, recall and f-measure scores were calculated to evaluate the quality of mappings generated. In addition, statistic analysis, namely two-tailed t-tests were carried out on the score generated by the SOCOM framework and the generic approach in order to validate the statistical significance of the experimental findings.

4. ACKNOWLEDGMENT

This research is partially supported by Science Foundation Ireland (Grant 07/CE/11142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie) at Trinity College Dublin.

5. REFERENCES

[1] Benjamins R. V., Contreras J., Corcho O., Gomez-Perez A. 2004. Six Challenges for the Semantic Web. *AIS SIGSEMIS Bulletin, Vol. 1, Iss. 1*, 2004.

[2] Euzenat J., Shvaiko P. 2007. Ontology Matching. *Springer-Verlag Berlin/Heidelberg.*

[3] Chang C., Lu W. 2002. The Translation of Agricultural Multilingual Thesaurus. In *Proceedings of the 3rd Asian Conference for Information Technology in Agriculture* (Beijing, China, October 26-28, 2002), 526-528.

[4] Shi C., Wang H. 2005. Research on Ontology-Driven Chinese-English Machine Translation. In *Proceedings of 2005 IEEE International Conference on Natural Language Processing & Knowledge Engineering* (Wuhan, China, October 30 – November 01, 2005), 426-430. DOI=10.1109/NLPKE.2005.1598775

[5] Guan J., Deng J, Qu Y. 2005. An Ontology-Driven Information Retrieval Mechanism for Semantic Information Portals. In *Proceedings of 1st International Conference on Semantic, Knowledge and Grid* (Beijing, China, November 27 - 29, 2005). SKG. IEEE Computer Society, Washington, DC, 63. DOI= http://dx.doi.org/10.1109/SKG.2005.42

[6] Zhang L., Wu G., Xu Y., Li W., Zhong Y. 2004. Multilingual Collection Retrieving Via Ontology Alignment. In Proceedings of the 7th International Conference on Asian Digital Libraries (Shanghai, China, December 13-17, 2004) LNCS 3334, 939-957. DOI=10.1007/978-3-540-30544-6_57

[7] Şah M., Wade V. 2010. Automatic Metadata Extraction from Multilingual Enterprise Content. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (Toronto, Canada, October 26-30, 2010), to appear.

[8] Steichen B., Wade V. 2010. Adaptive Retrieval and Composition of Socio-Semantic Content for Personalised Customer Care. In *Proceedings of International Workshop on Adaptation in Social and Semantic Web* (Big Island of Hawaii, USA, June 21, 2010), 1-10, ISSN 1613-0073.