The ESSOT System Goes Wild: an Easy Way For Translating Ontologies

Mihael Arcan[1], Mauro Dragoni[2], and Paul Buitelaar[1]

[1] Insight Centre for Data Analytics, National University of Ireland, Galway
[firstname.lastname]@insight-centre.org
[2] FBK- Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
dragoni@fbk.eu

Abstract. To enable knowledge access across languages, ontologies that are often represented only in English need to be translated into different languages. Since manual multilingual enhancement of domain-specific ontologies is very time consuming and expensive, smart solutions are required to facilitate the translation task for the language and domain experts. For this reason, we present ESSOT, an Expert Supporting System for Ontology Translation, which support experts in accomplishing the multilingual ontology management task.[3] Differently than the classic document translation, ontology label translation faces highly specific vocabulary and lack contextual information. Therefore, ESSOT takes advantage of the semantic information of the ontology for translation improvement of the ontology labels.

1 Introduction

Currently, most of the semantically structured data, i.e. ontologies or taxonomies, have labels stored in English only. Although, the increasing amount of ontologies offers an excellent opportunity to link this knowledge together, non-English users may encounter difficulties when using the ontological knowledge represented in English only [1]. Furthermore, applications in information retrieval or knowledge management, using monolingual ontologies are limited to the language in which the ontology labels are stored. Therefore, to make the ontological knowledge accessible beyond the language borders, these monolingual resources need to be enhanced with multilingual information [2]. For this reason, we engage a statistical machine translation (SMT) system, which takes into consideration the domain of the ontology to be translated. As ontologies may change over time, having in place an SMT system adaptable to an ontology can therefore be very beneficial. One of the main challenges in ontology translation are labels built out of only a few words, which do not often express enough semantic information to guide the SMT system to translate them into the targeted domain. This can be observed in domain-unadapted SMT systems, e.g. Google Translate,[4] where an ambiguous expression, like *vessel* stored in a medical ontology, is translated into a generic domain as *Schiff*[5] (en. *ship*) in German, but not into the targeted medical domain as *Gefäß*.

In this demo, we present our proposal for addressing the ontology translation task in a real-world settings. We will show the software modules composing the system,

[3] This demo paper is submitted as support of the accepted In-Use paper at ISWC 2016, in order to give the opportunity of showing more details on how the system works and how it has been used in different real-world settings. A read-only version, but with all functionalities, of the instance described in this paper is available at https://dkmtools.fbk.eu/moki/3_5/essot/
[4] https://translate.google.com/ [5] Translation performed on 06.07.2016

their functionalities and how they can be exploited as web services. Finally, we will provide information on how to engage the different components and how to prepare a local instance of the ESSOT platform.

2 System Implementation

Based on the lexical and semantic overlap with the ontology labels our proposed system identifies from a large English monolingual corpus the most relevant sentences containing the labels to be translated. The goal is to translate the ontology labels within the textual context of the targeted domain, rather than in isolation. For instance, with this selection approach, we aim to retain relevant sentences, where the English word *vessel* or *injection* belongs to the medical domain, but not to the technical domain.

Statistical Machine Translation For the translation approach, ESSOT engages the widely used Moses toolkit [3]. For a broader domain coverage of the SMT system we merged several parallel corpora, e.g. DGT (translation memories generated by the *Directorate-General for Translation*) [4], Europarl [5] and MultiUN corpus [6] among others, into one parallel data set necessary to train an SMT system. Due to the increasing amount of parallel data, ESSOT supports translations of English ontology labels into all (24) official languages of the European Union [7].

Query Expansion for Sentence Selection In order to improve the translation of ontology labels, we select from the concatenated corpus only those source sentences, which are most relevant to the labels to be translated [8]. The first criterion for relevance is the *n-gram overlap* between a label and a source sentence coming from the generic corpus. Once we obtain sentences with the targeted labels, we follow the idea of extending the semantic information of the labels using Word2Vec for computing distributed representations of words [9]. The technique is based on a neural network that analyses the textual data provided as input, in our experiment ontology labels and source sentences, and outputs a list of semantically related words. Each input string is vectorized and compared to other vectorized sets of words in a multi-dimensional vector space, which was trained with Word2Vec on Wikipedia articles.

To further improve the disambiguation of relevant sentences, the related words of the label are concatenated with the related words of its direct parent in the ontology hierarchy. Given a label and a source sentence from the generic corpus, related words are extracted from both of them, and used as entries of the vectors to calculate the cosine similarity. Finally, we translate the most similar source sentence with the targeted label and extract its translation once the translation approach is done.

User Facilities The ESSOT system integrates facilities supporting a collaborative translation of domain-specific ontologies in order to satisfy the requirements of the multilingual ontology enhancement from a user perspective. The system focuses on supporting two distinct experts groups: domain experts and language experts. Domain experts are in charge of the modelling aspect of ontologies (i.e. creation of concepts, individuals, properties, and the relationships between them). On the other hand, language experts are responsible for managing the labels associated with each entities by evaluating their correctness and, eventually, by providing a more fine-grained adaptation of the ontology with respect to the domain it represents.

The full set of facilities in ESSOT include: (i) *Experts Views*, which are in charge of presenting all information to experts in an effective manner; (ii) *Approval and Discussion* components, which are managing the collaborative workflow of entity editing

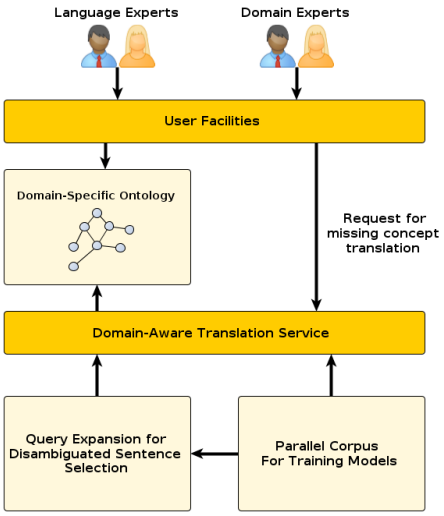

Fig. 1. ESSOT General Architecture.

by informing and providing experts with information necessary for understanding the status of each entity within the ontology; and (iii) the *Translator Connector* that is responsible of invoking the machine translation service, called OTTO [10], for providing a list of suggestions for translating the entity labels.

3 ESSOT in Action: What we Will Show During The Demo

The main part of the demo will be related to (i) the presentation of the general features of the platform, (ii) how a platform instance can be obtained and installed on local servers, and (iii) which are the mandatory parameters that have to be set for making the platform working on own servers. Furthermore, among the full set of features implemented into the ESSOT platform, our demo will focus on the ones described below.

Usability of The Tool. We will show how the process for translating an ontology works and how the user facilities can be used by the different type of experts in a collaborative way. In particular, we will focus on the "Approval Workflow" and how all the actors involved in the process of translating ontologies are notified about the multilingual enhancement of each entity. In addition to that, we will demonstrate how the underlying machine translation components suggests candidate translations to the experts and how such suggestions can be selected for their inclusion the ontology.

Plug-and-Play of Translation Models. A first more technical demonstration is related to the plug-and-play facility of the platform for creating and connecting different machine translation models and/or services. We will show how developers can configure the platform in simple steps by connecting it to machine translation models stored locally or to external translation services (i.e. Microsoft Bing).

Usage of ESSOT as Web Service. Finally, the machine translation service integrated into the ESSOT platform can be queried from third party applications by exploiting the available RESTful interface.[6] We will show how the service works, which are the expected inputs and the structure of the output.

[6] `http://server1.nlp.insight-centre.org/otto/rest_service.html`

4 Conclusion

This paper is aimed at showing ESSOT for multilingual management of semantically structured data, i.e. ontologies or taxonomies. The system is based on an approach to identify the most relevant source sentences from a large generic parallel corpus, giving the possibility to automatically translate highly specific ontology labels in context without particular in-domain parallel data. The demonstrated approach reduces the ambiguity of expressions in the selected sentences, which consequently generates better translations of ontology labels. As an ongoing work, we further focus on improving the extraction of the lexical knowledge stored in ontologies. Additionally, we plan to enable knowledge enrichment for existing multilingual ontologies.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. Gómez-Pérez, A., Vila-Suero, D., Montiel-Ponsoda, E., Gracia, J., Aguado-de Cea, G.: Guidelines for multilingual linked data. In: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, ACM (2013)
2. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web **11** (2012)
3. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Stroudsburg, PA, USA (2007)
4. Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., Gilbro, S.: An overview of the european union's highly multilingual parallel corpora. Language Resources and Evaluation **48**(4) (2014) 679–707
5. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit, AAMT (2005)
6. Eisele, A., Chen, Y.: Multiun: A multilingual corpus from united nation documents. In Tapias, D., Rosner, M., Piperidis, S., Odjik, J., Mariani, J., Maegaard, B., Choukri, K., Chair), N.C.C., eds.: Proceedings of the Seventh conference on International Language Resources and Evaluation, European Language Resources Association (ELRA) (5 2010) 2868–2872
7. Arcan, M., Dragoni, M., Buitelaar, P.: ESSOT: an expert supporting system for ontology translation. In Métais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S., eds.: Natural Language Processing and Information Systems - 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings. Volume 9612 of Lecture Notes in Computer Science., Springer (2016) 60–73
8. Arcan, M., Turchi, M., Buitelaar, P.: Knowledge portability with semantic expansion of ontology labels. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, Beijing, China (July 2015)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. ICLR Workshop (2013)
10. Arcan, M., Asooja, K., Ziad, H., Buitelaar, P.: Otto – ontology translation system. In: ISWC 2015 Posters & Demonstrations Track. Volume 1486., Bethlehem (PA), USA (2015)