The ESSOT system goes wild: an easy way for translating ontologies

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. This activity is time consuming, therefore, smart solutions are required for facilitating the translation task from 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]. Rather a different task 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 to improve the translation of 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].

Since manual multilingual enhancement of domain-specific ontologies is very time consuming and expensive, we engage a domain-aware statistical machine translation (SMT) system to automatically translate the ontology labels. As ontologies may change over time; having in place an SMT system adaptable to an ontology can therefore be very beneficial. Nevertheless, the quality of the SMT generated translations relies strongly on the translation model learned from the information stored in parallel corpora. In most cases, the inference of translation candidates cannot always be learned

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

accurately when specific vocabulary, like ontology labels, appears infrequent in a parallel corpus. Additionally, ambiguous labels built out of only a few words do not often express enough semantic information to guide the SMT system to translate a label 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*.

2 Related Work

The task of ontology translation involves generating an appropriate translation for the lexical layer, i.e. labels stored in the ontology. Most of the previous related work focused on accessing existing multilingual lexical resources, like EuroWordNet or IATE [3, 4]. Their work focused on the identification of the lexical overlap between the ontology and the multilingual resources, which guarantees a high precision but a low recall. Consequently, external translation services like BabelFish, SDL FreeTranslation tool or Google Translate were used to overcome this issue [5, 6]. Additionally, [5] and [7] performed ontology label disambiguation, where the ontology structure was used to annotate the labels with their semantic senses. Differently to the aforementioned approaches, which rely on external knowledge or services, we focus on how to gain adequate translations with a domain-aware SMT system, which is supported by the ontology hierarchy. Our back-end approach is then integrated with a set user facilities and implemented into the ESSOT platform for supporting experts in translating ontologies.

3 System Implementation

Based on the lexical and semantic overlap with the ontology labels ESSOT[6] identifies, from a large set of parallel corpora, the most relevant source 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 Moses toolkit [**?**]. To have a broader domain coverage of the generic parallel dataset necessary for training the SMT system, we merged the JRC-Acquis 3.0 [8], Europarl v7 [9] and OpenSubtitles2013 [10], thus obtaining a training corpus of 8.5M parallel sentences for English-German, 18.9M for English-Italian and 33.6M for the English-Spanish translation directions. To train OTTO for the (under-resourced) English-Irish translation direction, we collected around 723K parallel sentences from various parallel corpora, like DGT (DG Translation at the European Commission), EUbookshop or KDE4, from the OPUS webpage.[7]

[4] https://translate.google.com/ [5] Translation performed on 06.07.2016

[6] http://server1.nlp.insight-centre.org/otto/ [7] http://opus.lingfil.uu.se/

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. The first criterion for relevance is the *n-gram overlap* between a label and a source sentence coming from the generic corpus. Due to the specificity of the ontology labels, just an *n-gram overlap* approach is not sufficient to select all the useful sentences. For this reason, we follow the idea of extending the semantic information of the labels using Word2Vec[8] for computing distributed representations of words [11]. 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 [12]. 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 the Wikipedia articles.[9]

To further improve the disambiguation of short labels, 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 and their weights are extracted from both of them, and used as entries of the vectors to calculate the cosine similarity. Finally, the most similar source sentence and the label should share the largest number of related words.

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

The full set of facilities include: (i) "Experts Views": these modules are in charge of presenting all information to experts in an effective manner; (ii) "Approval and Discussion" components: they are in charge of managing the collaborative workflow of entity editing by informing and providing experts with information that are 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 for providing a list of suggestions for translating entity labels.

4 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.

[8] https://code.google.com/p/word2vec/ [9] Wikipedia dump id enwiki-20141106

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 we will show how all the actors involved in the process of translating ontologies (and of revising such translations) are notified about what occurs on each entity every time. Then, we will show 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. We will show how the service works, which are the expected inputs and the structure of the output.

5 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. Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., Gómez-Pérez, A.: A note on ontology localization. Appl. Ontol. **5**(2) (2010) 127–137
4. Declerck, T., Pérez, A.G., Vela, O., Gantner, Z., Manzano, D.: Multilingual lexical semantic resources for ontology translation. In: In Proceedings of the 5th International Conference on Language Resources and Evaluation, Saarbrücken, Germany (2006)
5. Espinoza, M., Montiel-Ponsoda, E., Gómez-Pérez, A.: Ontology localization. In: Proceedings of the Fifth International Conference on Knowledge Capture, NY, USA (2009)
6. Fu, B., Brennan, R., O'Sullivan, D.: Cross-lingual ontology mapping - an investigation of the impact of machine translation. In Gómez-Pérez, A., Yu, Y., Ding, Y., eds.: ASWC. Volume 5926 of Lecture Notes in Computer Science., Springer (2009)
7. McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de Cea, G., Cimiano, P.: Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In: Fifth workshop on Syntax, Structure and Semantics in Statistical Translation (SSST-5). (2011)
8. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). (2006)
9. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit, AAMT (2005)
10. Tiedemann, J.: Parallel data, tools and interfaces in opus. In Chair), N.C.C., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eight International Conference on Language Resources and Evaluation, Istanbul, Turkey (may 2012)
11. 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)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. ICLR Workshop (2013)
13. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the semantic web with lemon. The Semantic Web: Research and Applications (2011)