



## OTTO - ontology translation system

Title	OTTO - ontology translation system
Author(s)	Arcan, Mihael;Asooja, Kartik;Ziad, Housam;Buitelaar, Paul
Publication Date	2015-08-11
Publisher	CEUR-WS.org

# OTTO – Ontology Translation System

Mihael Arcan    Kartik Assoja    Housam Ziad    Paul Buitelaar  
Insight Centre for Data Analytics @ NUI Galway, Ireland  
{firstname.lastname}@insight-centre.org

**Abstract.** To enable knowledge access across languages, ontologies that are often represented only in English, need to be translated into different languages. For this reason, we present OTTO, an OnTology TranslatiOn System, which enhances ontologies with multilingual information. Rather a different task than the classic document translation, ontology label translation faces highly specific vocabulary and lack contextual information. Therefore, OTTO 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 [6]. 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 [7].

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 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,<sup>1</sup> where an ambiguous expression, like *vessel* stored in a medical ontology, is translated into a generic domain as *Schiff*<sup>2</sup> (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

<sup>1</sup> <https://translate.google.com/>

<sup>2</sup> Translation performed on 25.06.2015

work focused on accessing existing multilingual lexical resources, like EuroWordNet or IATE [2, 3]. 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 [4, 5]. Additionally, [4] and [10] 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.

### 3 System Implementation

Based on the lexical and semantic overlap with the ontology labels, the OnTology TranslatiOn System – OTTO<sup>3</sup> 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, OTTO engages the Moses toolkit [9]. To have a broader domain coverage of the generic parallel dataset necessary for training the SMT system, we merged the JRC-Acquis 3.0 [13], Europarl v7 [8] and OpenSubtitles2013 [14], 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.<sup>4</sup>

**Relevant 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<sup>5</sup> for computing distributed representations of words [1]. 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.<sup>6</sup>

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,

<sup>3</sup> <http://server1.nlp.insight-centre.org/otto/>

<sup>4</sup> <http://opus.lingfil.uu.se/>

<sup>5</sup> <https://code.google.com/p/word2vec/>

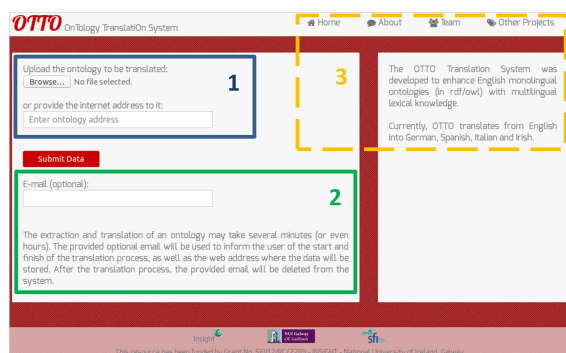
<sup>6</sup> Wikipedia dump id enwiki-20141106

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.

## 4 OTTO Demo

OTTO takes as input an ontology represented in OWL or RDF and extracts the labels stored in it. To improve the translations of the labels stored in the ontology, the most relevant sentences, which contain the labels, are obtained from the concatenated generic corpus. Once the labels are identified in the context of the relevant sentences, OTTO engages Moses and translates the labels within the context into German, Italian, Spanish and Irish. After the translation process is done, the translated labels are identified in the relevant target sentences.

Since the translation of the extracted labels may take several minutes (or even hours), the OTTO user can provide an optional e-mail,<sup>7</sup> which is used to inform the user about the completion of the translation process, as well as the web address where the provided data will be stored. Without this information, the address of the stored data is given through the OTTO interface (Figure 1).



**Fig. 1.** The graphical interface of OTTO, with the input options (1), optional e-mail of the user (2) and additional information about the system (3).

In the last step, the translated labels are represented in different formats, for example in a HTML table and CSV file to allow a better visualisation. Furthermore, the multilingual information is injected into the original monolingual ontology and represented as a multilingual ontology as well as in *lemon*<sup>8</sup> [11], a model for linking linguistic information with ontologies.

## 5 Conclusion

This paper is aimed at showing OTTO, an OnTology TranslatiOn System for multilingual enrichment 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

<sup>7</sup> The provided e-mail is stored as a variable and is deleted after the process finishes.

<sup>8</sup> <http://lemon-model.net/>

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. 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 and the 7th International Joint Conference on Natural Language Processing. ACL, Beijing, China (2015)
2. Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., Gómez-Pérez, A.: A note on ontology localization. *Appl. Ontol.* 5(2), 127–137 (Apr 2010)
3. Declerck, T., Pérez, A.G., Vela, O., Gantner, Z., Manzano, D., D-Saarbrücken: Multilingual lexical semantic resources for ontology translation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (2006)
4. Espinoza, M., Montiel-Ponsoda, E., Gómez-Pérez, A.: Ontology localization. In: Proceedings of the Fifth International Conference on Knowledge Capture. K-CAP '09, ACM, New York, NY, USA (2009)
5. 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. Lecture Notes in Computer Science, vol. 5926. Springer (2009)
6. 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)
7. 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)
8. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit. AAMT (2005)
9. 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)
10. 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)
11. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the semantic web with lemon. *The Semantic Web: Research and Applications* (2011)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781 (2013)
13. 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)
14. 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)