Web Semantics: Science, Services and Agents on the World Wide Web 36 (2016) 23–31

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Domain adaptation for ontology localization

John P. McCrae [a,b,*], Mihael Arcan [b], Kartik Asooja [b,c], Jorge Gracia [c], Paul Buitelaar [b], Philipp Cimiano [a]

[a] Cognitive Interaction Technology, Center of Excellence, Universität Bielefeld, Inspiration 1, 33615 Bielefeld, Germany
[b] Insight Centre for Data Analytics, National University of Ireland, Galway, IDA Business Park, Galway, Ireland
[c] Ontology Engineering Group, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Boadilla del Monte, Spain

HIGHLIGHTS

- Detailed description of an architecture and methodology for machine translation of ontologies.
- Methodology for extracting domain terminology from several resources.
- Statistical methods for the domain adaptation of machine translation systems according to ontologies.
- Detailed evaluation showing improvement in translation quality for a number of ontologies.

ARTICLE INFO

Article history:
Received 27 March 2014
Received in revised form
22 December 2015
Accepted 22 December 2015
Available online 30 December 2015

Keywords:
Ontology localization
Statistical machine translation
Domain adaptation

ABSTRACT

Ontology localization is the task of adapting an ontology to a different cultural context, and has been identified as an important task in the context of the Multilingual Semantic Web vision. The key task in ontology localization is translating the lexical layer of an ontology, i.e., its labels, into some foreign language. For this task, we hypothesize that the translation quality can be improved by adapting a machine translation system to the domain of the ontology. To this end, we build on the success of existing statistical machine translation (SMT) approaches, and investigate the impact of different domain adaptation techniques on the task. In particular, we investigate three techniques: (i) enriching a phrase table by domain-specific translation candidates acquired from existing Web resources, (ii) relying on Explicit Semantic Analysis as an additional technique for scoring a certain translation of a given source phrase, as well as (iii) adaptation of the language model by means of weighting *n*-grams with scores obtained from topic modelling. We present in detail the impact of each of these three techniques on the task of translating ontology labels. We show that these techniques have a generally positive effect on the quality of translation of the ontology and that, in combination, they provide a significant improvement in quality.

1. Introduction

The vision of a Multilingual Web of Data in which knowledge is represented in a language-independent fashion and users can access this knowledge in their own language, has attracted the attention of research efforts in the area of the Semantic Web recently [1,2]. In fact, the Web of Data is moving from a monolingual landscape (in English) towards hosting an increasing amount of multilingual content. For instance, the number of multilingual RDF datasets on the Web doubled from January 2012 to December 2012 [3]. However, realizing the Multilingual Web vision according to which users can access semantic information in any natural language requires the localization of the vocabularies that the information is described with. The task of translating ontological vocabularies into other languages is thus at the core of the Multilingual Semantic Web Vision, and high-quality translation approaches are required [4]. This task involves the translation of ontology labels and, as manual translation of existing vocabularies is a time-intensive and costly process, automatic techniques, such as the one proposed in this paper, are needed. Furthermore, these labels are frequently only fragments of text, instead of

* Corresponding author at: Insight Centre for Data Analytics, National University of Ireland, Galway, IDA Business Park, Galway, Ireland.
E-mail addresses: john@mccr.ae (J.P. McCrae), mihael.arcan@insight-centre.org (M. Arcan), kartik.asooja@insight-centre.org (K. Asooja), jgracia@fi.upm.es (J. Gracia), paul.buitelaar@insight-centre.org (P. Buitelaar), cimiano@cit-ec.uni-bielefeld.de (P. Cimiano).

http://dx.doi.org/10.1016/j.websem.2015.12.001
1570-8268/

full sentences as typically handled by state-of-the-art machine translation systems. Indeed, off-the-shelf machine translation systems are not designed to translate the short labels that typically occur as labels of ontology elements in SW ontologies, but typically require more context (i.e., a full sentence) to yield satisfactory translation results. Our goal is to develop methods that factor in the ontological context of a label into the translation task, making standard SMT systems also applicable to the task of localizing ontologies. With ontological context we refer to the semantic neighbourhood of a given concept within an ontology, in particular the neighbours in the graph occurring within a fixed distance from the ontology element the label of which is to be translated. In this line, in this paper we investigate the impact of multiple domain adaptation techniques with respect to the task of ontology localization. In this paper we handle smaller ontologies for which the context of a label can be considered to be the whole ontology. However, for very large ontologies such as DBpedia [5], techniques to identify the more immediate context should be applied.

Our approach to domain adaptation takes three complementary paths as extension to a state-of-the-art and off-the-shelf statistical machine translation (SMT) system such as Moses [6], which relies on a probabilistic model learned from a parallel corpus coupled with a monolingual language model acquired from a larger monolingual corpus to score the plausibility of a translation.

Firstly, we consider enriching the phrase table used by the machine translation system by translation candidates that are specific to the domain. In this case, we use the labels in the ontology to bootstrap this process and extract translation candidates from Wikipedia and other resources.

Our second approach involves the direct incorporation of the semantic context of the ontology label into the translation model. This is achieved by incorporating a feature which describes how semantically similar a potential translation is to the ontology, by means of a score computed by Cross-Lingual Explicit Semantic Analysis (CL-ESA) [7,8].

Finally, our third approach consists in adjusting the translation model itself in response to the domain of the translation. We achieve this by means of updating the language model with new probabilities that are learnt by weighting each document in the corpus individually by way of its similarity to the ontology as a whole.

We quantify the impact of all these domain adaptation techniques on the task of ontology localization using a state-of-the-art statistical machine translation system as baseline [6]. We show that all individual domain adaptation techniques lead to some improvement. The impact actually comes from using all domain adaptation techniques in combination, which yields an improvement of up to 30 points in BLEU score [9] according to our experiments for the financial domain.

The paper is structured as follows: Section 2 discusses the framework and architecture of the system we propose and which builds on a state-of-the-art statistical machine translation (SMT) framework. In Sections 3–5 we present in more detail the three domain adaptation techniques examined. Section 6 reports on our experiments on 2 ontologies: the IFRS ontology and a public service ontology that were used as use cases in the FP7 Monnet Project. We describe the datasets we use in more detail as well as the evaluation metrics used. We first present and discuss the results of the single components with respect to a baseline system and then move to discuss results of applying the mentioned domain adaptation techniques in combination. Before concluding, we discuss some related work in Section 7.

2. Framework and architecture

In this section we briefly review the traditional statistical MT approach and give an overview of our proposed architecture for ontology translation.

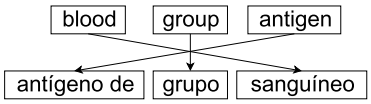

Fig. 1. An example of constructing a translation by phrase-based statistical machine translation.

2.1. Statistical machine translation

We base our approach on the statistical approach to machine translation [10], where we wish to find the translation that maximizes some function such that the best translation, **t**, of a foreign label, **f**, is given by a log-linear model combining some set of features $\{\phi_i(\mathbf{t}|\mathbf{f})\}$:

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \prod_i \exp(w_i\, \phi(\mathbf{t}|\mathbf{f})) = \arg\max_{\mathbf{t}} \sum_i w_i\, \phi_i(\mathbf{t}|\mathbf{f}). \quad (1)$$

The translation that maximizes the score of the log-linear model is obtained by searching in the space of possible translations via a so called *decoder*. The decoder is essentially a search procedure that computes the sentence in the target language that maximizes the above score given some statistical translation model induced from the training data. Hereby, it is assumed that both **t** and **f** are segmented into a number of phrases, t_i and f_i, and that we have a *phrase table* consisting of pairs of translations $\{(t_i, f_i)\}$. A *candidate* translation is one such that every phrase in **f** can be paired with a phrase in **t** and this pair occurs in the phrase table.[1] In this model, we take the standard set of features as used in the Moses system [6]. These are given as follows:

- The logarithm of the probability, $p(t_i|f_i)$, that is the probability that f_i is translated as t_i.
- The logarithm of the lexical weighting of t_i given f_i [12] summed over all phrases.
- The logarithm of the probability, $p(f_i|t_i)$, that is the probability that t_i is translated as f_i.
- The logarithm of the lexical weighting of f_i given t_i summed over all phrases.
- The number of phrases used in the segmentation.
- The logarithm of the language model probability, a score of the plausibility of the translation according to a statistical n-gram model of the target language.
- The number of unknown phrases used in the translation.
- The distortion model. For each pair (f_i, t_i), the feature indicates the number of words this pair has been moved away from each other.

For example, in Fig. 1 we see the translation of the English ontology label "blood group antigen" into a Spanish label "antígeno de grupo sanguíneo". The scores for the translation would be given by the scores for each feature for the aligned phrases, e.g., "antigen" and "antígeno de". The best translation is then found by a heuristic beam or stack search.

[1] In order to deal with unknown words not observed during training, unknown phrases of length 1 are assumed to translate to themselves. We note that it would be possible to apply a transliteration method in this case [11], but we do not consider this in the context of this work.