

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*

E-mail: john@mccr.ae, cimiano@cit-ec.uni-bielefeld.de

^b*Insight Centre for Data Analytics, National University of Ireland, Galway, IDA Business Park, Galway, Ireland*

E-mail: {mihael.arcan,kartik.asooja,paul.buitelaar}@insight-centre.org

^c*Ontology Engineering Group, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Boadilla del Monte, Spain*

E-mail: jgracia@fi.upm.es

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 hypothesise 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.

Keywords: ontology localization, statistical machine translation, domain adaptation

^{*}Corresponding Author

Email address: john@mccr.ae (John P. McCrae)

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 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 neighborhood 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
50 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
55 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
60 architecture of the system we propose and which builds on a state-of-the-art statistical machine translation (SMT) framework. In Section 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
65 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.

70 **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.

2.1. Statistical machine translation

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

$$\begin{aligned}\hat{\mathbf{t}} &= \arg \max_{\mathbf{t}} \prod_i \exp(w_i \phi_i(\mathbf{t}|\mathbf{f})) \\ &= \arg \max_{\mathbf{t}} \sum_i w_i \phi_i(\mathbf{t}|\mathbf{f})\end{aligned}\tag{1}$$

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
80 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 \mathbf{t} and \mathbf{f} are segmented into a number of phrases, \mathbf{t}_i and \mathbf{f}_i , and that we have a *phrase table* consisting of pairs of translations $\{(\mathbf{t}_i, \mathbf{f}_i)\}$. A *candidate* translation is one


blood group antigen

antígeno de grupo sanguíneo

Figure 1: An example of constructing a translation by phrase-based statistical machine translation

such that every phrase in \mathbf{f} can be paired with a phrase in \mathbf{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, $p(\mathbf{t} | \mathbf{f})$, that is the probability that \mathbf{f} is translated as \mathbf{t}
- The logarithm of the lexical weighting of \mathbf{t} given \mathbf{f} [12] summed over all phrases
- The logarithm of the probability, $p(\mathbf{f} | \mathbf{t})$, that is the probability that \mathbf{t} is translated as \mathbf{f}
- The logarithm of the lexical weighting of \mathbf{f} given \mathbf{t} 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 (\mathbf{f}, \mathbf{t}) , the feature indicates the number of words this pair has been moved away from each other.

For example, in Figure 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.

2.2. Our architecture

Our architecture for domain-adapted ontology translation, illustrated in Figure 2, consists of the following components:

¹In 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.

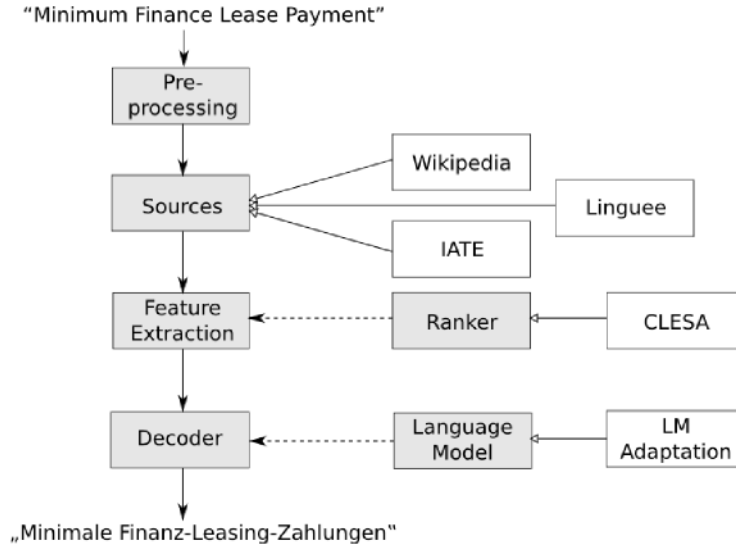


Figure 2: The architecture of our machine translation system

Pre-processing. This performs the segmentation of the input label, first by tokenizing the label and then by finding all relevant subsequences of this label. In practice, we simply use an exhaustive method that returns all subsequences of the label.

Phrase table augmentation (Sources). As phrase tables are typically very large (often of the order of several GB), loading the phrase table can present a significant performance bottleneck to the machine translation system. Thus, we store all phrase tables in a database and load all potentially relevant translations into memory for each translation. In this stage we also include any translations that come from other domain-specific sources.

Domain-Specific Feature Extraction. These extend the baseline translation model as described above by including additional features for translation. These come from raw scores from a ranker. The feature extractor’s job is to reconcile this with the scores from various sources. In particular, we introduce a ranker based on Cross-Lingual Explicit Semantic Analysis (CLESA).

Domain Language Model Adaptation. We also include a language model, which can be built either based on the input domain ontology, or reuses an existing domain ontology model.

Decoder. The decoder combines all the phrase tables augmented with the additional features from the rankers along with the language model to find

an (approximately) optimal translation.

The architecture is implemented in Java in a modular fashion and is available for download at <http://github.com/monnetproject/translation>.
The overall architecture is depicted in Figure 2.

3. Domain-targeted Phrase Table Augmentation

3.1. Automatic enrichment of the phrase table

The first part of our domain adaptation method for ontology translation consists in the extraction of domain-specific translations from different resources. These additional translation candidates are added to the phrase table of the SMT system. A phrase table contains pairs of translations from source to target language as well as the first four features of the translation model for this phrase. An excerpt of a phrase table for ‘antigen’ from English to Spanish would look as follows:

English	Spanish	p(ti fi)	lex(ti fi)	p(fi ti)	lex(fi ti)
antigen	antigen	0.821	0.731	0.043	0.033
antigen	antigénico	0.214	0.088	0.004	0.002
antigen	antigénicos	0.333	0.125	0.001	0.001
antigen	antídoto	0.002	0.001	0.002	0.001
antigen	antígeno de la	0.250	0.250	0.003	0.004
antigen	antígenos	0.051	0.167	0.020	0.046

We used DBpedia to extract the relevant titles and their equivalents in other languages from domain-specific Wikipedia articles. Further, the labels of the ontology are used to query the Linguee Web service² to yield parallel text in which the labels are translated to other languages with the corresponding linguistic context.

3.1.1. Domain terminology from Wikipedia

For the domain-specific terminology extraction we used the datasets provided by the DBpedia project [5].
In order to improve translations of highly domain-specific vocabulary, the method described here derives a bilingual domain-specific translation lexicon from the DBpedia datasets (version 3.8)³.
In particular, the method exploits the *Articles Categories* dataset, which links Wikipedia titles to categories using the SKOS vocabulary [13]. In order to extend the vocabulary of a specific domain, the method uses the *Wikipedia Pagelinks* dataset, which contains the internal links between Wikipedia articles

²<http://www.linguee.com/>

³<http://wiki.dbpedia.org/Downloads38>.

as well as the *Inter-Language Links* dataset, which contains interlanguage links between many Wikimedia projects. From these datasets, the method extracts the following information: the Wikipedia article titles, the variants and the translations of article titles, and the categories associated with these articles. With this information, we build a cross-lingual terminological lexicon, exploiting two approaches:

- a) domain detection of the ontology (bottom-up approach);
- b) extraction of cross-lingual terminology (top-down approach).

In our first step, the method uses the DBpedia knowledge base to determine the domain of the ontology. The bottom-up approach consists of representing the domain by the most frequent categories associated with the vocabulary to be translated. For this approach, the labels, which are extracted from the ontology, as well as all token subsequences (i.e., n -grams) are used to query the DBpedia knowledge base. If a label or n -gram exactly matches an article title, all categories associated with this article are collected. This results in a list of categories together with a number of labels or n -grams which support that category. We refer to the number of distinct n -grams that generate the category simply as the *category frequency*. An example is given in Table 1.

Frequency	Wikipedia Category Name
8	Generally Accepted Accounting Principles
4	Debt
4	Accounting terminology
4	Economics terminology
...	...
1	Political science terms
1	Physical punishments

Table 1: Collected Wikipedia categories (prefinalCategoryList) based on the extracted financial (sub-)labels

After collecting all categories, categories which are not relevant for the domain are filtered out. This is performed heuristically by eliminating those categories which have a frequency lower than or equal to the mean frequency. When calculating the mean frequency, categories with a frequency of 1 are ignored to prevent the mean from being artificially low.

That is, we only consider categories C for which the category frequency f_c fulfils the following:

$$f_c > \frac{\sum_{C \in \mathcal{C}, f_c > 1} f_c}{|\{C \in \mathcal{C} : f_c > 1\}|}$$

In the next step, the list of collected categories is further extended by exploiting Wikipedia links of each article the title of which is equivalent to a label

Frequency	Wikipedia Category	Name
95	Economics terminology	
62	Generally Accepted Accounting Principles	
61	Macroeconomics	
55	Accounting terminology	
47	Finance	
44	Economic theories	
42	International trade	
...		

Table 2: Most frequent Categories based on the German GAAP labels and their Pagelinks (finalCategoryList)

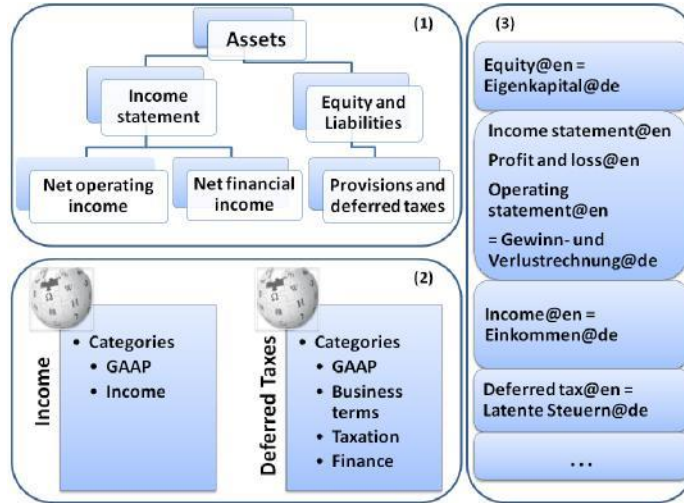


Figure 3: Steps of extraction Wikipedia titles and its translations

or n -gram of a label in the ontology. For each of these articles, the categories
of outgoing links are selected and their frequencies are recalculated and
filtered as described previously.

For all the categories extracted as described above, as shown for example in
Table 2, all translations from/to German, Spanish or Dutch from Wikipedia
articles assigned to these categories are extracted and stored in the bilingual
lexicon. This extraction process is shown in Figure 3.

3.1.2. Domain-specific parallel resource acquisition from Linguee

In order to enrich the phrase table with additional domain-specific candidates,
we built a new parallel corpus based on the taxonomy vocabulary that we want to
translate. For this we used Linguee, a combination of a dictionary and a search
engine which indexes words and expressions from around 100 mil-
lion bilingual texts. Linguee search results display example sentences that show

how the expression searched for has been translated in context.

The Linguee bilingual dataset represents a very large collection of manually translated sentences in English, German, Spanish, French, Italian and
205 Portuguese collected from the Web, in particular from multilingual websites of companies, organisations, universities and other sources, including EU documents and patent specifications. Recently Japanese, Chinese, Polish and Dutch bilingual data was added.

For generating a domain-specific parallel resource, the Linguee search engine
210 was queried with labels extracted from our ontologies. For each query, Linguee provides aligned parallel sentences through their web service. We extracted the output and stored the source and target sentences separately, which were finally used to build domain-specific translation models.

The domain-specific translation models and language models were used in 215 the ontology translation process independently as well as in combination with other domain adaptation techniques (see Section 6.5).

3.1.3. Domain terminology from IATE

In addition, our method allows for the incorporation of further terminological resources. In particular, we included a very large multilingual term base
220 called Inter-Active Terminology for Europe (IATE) ⁴ as an additional source of translations. IATE is the European Union (EU) inter-institutional terminology database. It contains all the existing terminology databases of the EU's translation services. However, the multilingual term-bases may contain several possible translations for a single term in different domains, and disambiguation
225 is required while adding the translations from such a multi-lingual term base. For this purpose, we rely on Cross Lingual Explicit Semantic Analysis (CLESA) as an additional semantic ranker to the translation architecture.

4. Domain-Specific Feature Extraction

4.1. Cross-lingual Explicit Semantic Analysis

230 Explicit Semantic Analysis (ESA) was introduced by Gabrilovich and Markovitch [7], and supports the comparison of texts with respect to their semantic similarity by indexing the texts in a space defined by explicit concepts. In contrast, other techniques such as Latent Semantic Analysis [14] and Latent Dirichlet Allocation [15] build unsupervised concepts by the correlations of the
235 terms in the data. ESA is an algebraic model in which the text is represented by a vector of the explicit concepts. The magnitude of each dimension in the vector is the associativity weight of the text to that explicit concept/dimension. In order to quantify this association, the textual content related to the explicit concept/dimension is utilized. This weight can be calculated by considering
240 different methods, for instance, we utilized the Lucene scoring function⁵. A

⁴<http://iate.europa.eu/>

⁵http://lucene.apache.org/core/3_6_2/scoring.html

possible way of defining concepts in ESA is by means of using the Wikipedia titles as dimensions of the model and the corresponding articles for calculating the associativity weight [7], thus taking advantage of the vast coverage of the community-developed Wikipedia.

245 A compelling characteristic of Wikipedia is the large collective knowledge available in multiple languages, which facilitates an extension of the original ESA model to accommodate multiple languages called Cross-lingual Explicit Semantic Analysis (CLESA) [8]. The articles in Wikipedia are linked together across languages and this cross-lingual link structure can provide a mapping of
250 a vector in one language to the other. Thus, Wikipedia provides the comparable corpus in different languages, which is required by CLESA.

To illustrate CLESA, let's take two ontology labels, f in the source language and t in the target language. As a first step, a concept vector for f is created using the Wikipedia corpus in the source language based on the tokens used in
255 the label. Similarly, the concept vector for t is created in the target language. Then, one of the concept vectors can be converted to the other language by using the cross-lingual mappings provided by Wikipedia. After obtaining both of the concept vectors in one language, the relatedness of the ontology labels f and t can be calculated by using cosine product, as in monolingual ESA.

260 MT systems implicitly use the local context for a better lexical choice during the translation [16]. Accordingly, it is natural to assume that a focused Word Sense Disambiguation (WSD) system integrated into an SMT system might produce better translations. We follow an approach in which we directly incorporate Word Sense Disambiguation into the SMT system as a multi-word phrasal
265 lexical disambiguation system [17]. The WSD probability score calculated by using CLESA is added as an additional feature in the log-linear translation model. The CLESA based score would depend on the ontology context into which the label is embedded, which could thus be exploited when determining the appropriate translation of the label. The CLESA score is then included as
270 an extra feature in equation 1. The score is applied to calculate the semantic similarity between the ontological context of the ontology label as a bag-of-word vector containing labels of neighbouring classes and the translation candidates considered by the SMT system.

5. Domain Language Modelling Adaptation

275 The language model calculates the probability of a given sentence in the target language by means of an n -gram approximation to the probability of translation. That is

$$H_{p(w_1 \dots w_m)} = \prod_{i=1, \dots, m} p(w_i | w_{i-1} \dots w_{\max(i-n, 1)})$$

These probabilities can be estimated simply by counting the occurrences in the corpus of an n -gram in order to obtain an unnormalized count $c(w_1 \dots w_n)$.
280 The conditional probability can then be obtained as usual by:

$$p(w_n|w_1 \dots w_{n-1}) = \frac{c(w_1 \dots w_n)}{c(w_1 \dots w_{n-1} \square)}$$

Where \square indicates any word. Furthermore, we can consider that these counts are obtained by summing over a corpus consisting of a set of documents $D = \{d_1, \dots, d_n\}$. The count can thus be expressed as follows:

$$\sum_{d_i \in D} c(w_1 \dots w_n) = \sum_{d_i \in D} c(w_1 \dots w_n \square d_i)$$

For the purpose of domain adaptation of the language model, we estimate the relevance of each document to our input ontology O by means of a similarity metric $so(d_i)$. In this way, we can obtain a modified count as follows

$$\sum_{d_i \in D} \tilde{c}(w_1 \dots w_n) = \sum_{d_i \in D} so(d_i) c(w_1 \dots w_n \square d_i)$$

Thus, our language model obtains the probability with the modified formula:⁶

$$p(w_n|w_1 \dots w_{n-1}) = \frac{\tilde{c}(w_1 \dots w_n)}{\tilde{c}(w_1 \dots w_{n-1} \square)}$$

After experimenting with a number of metrics, we found that the most effective measure of similarity between ontology and document is given by the cosine similarity of the word frequency vector of the ontology and the document:

$$so(d_i) = \cos(\mathbf{tfo}, \mathbf{tf}_{d_i}) = \frac{\mathbf{tfo}^T \mathbf{tf}_{d_i}}{\|\mathbf{tfo}\| \|\mathbf{tf}_{d_i}\|}$$

where \mathbf{tfo} represents the normalized word frequency of all words occurring in labels of entities in the ontology, and \mathbf{tf}_{d_i} represents the normalized word frequency of all words occurring in $\{w \square d_i\}$ the document d_i , i.e.,

$$\mathbf{tf}_{d_i}(w) = \frac{1}{|\{w \square d_i\}|} \mathbb{E}_i \{w \square d_i\}$$

If we assume that we have documents that are aligned across languages, for example Wikipedia articles aligned across topics, then we can make a further assumption that the similarity so should be approximately equal regardless of which language the document is in. Building on this assumption, we calculate the similarity of each document to the ontology in the foreign language, but use these scores to generate n -gram counts in the translation language. The above mentioned unmodified counts are finally smoothed using Modified Kneser-Ney smoothing [18].

⁶This method is labelled 'LM' in our results

Ontology	Labels
IFRS 2009	2,757
DE-GAAP	2,782
LAG	196
RB	1,449
HB	857

Table 3: The size of the ontologies used in our evaluations

6. Experiments and Results

In this section, we discuss the evaluation of the system with domain-specific ontologies. We first introduce the test data and evaluation metrics used in our experimentation. Then we describe the experiment and discuss the obtained results.

6.1. Datasets

We applied our ontology translation system to two domains, firstly a financial domain, where the ontologies are expressed in the XBRL standard [19], and secondly to a set of ontologies describing public services, provided by partners in the Monnet project. For the financial domain, we select two ontologies corresponding to the 2009 International Finance Reporting Standard (IFRS 2009) and the German Generally Accepted Accounting Principles (DE-GAAP). We use DE-GAAP as a development ontology to apply our domain adaptation and tested on IFRS in the only common language pair, i.e. English-German. The public service ontologies are smaller and we refer to them only by abbreviation (LAG, RB and HB).

6.2. Evaluation Methodology

The evaluation methodology was as follows: when translating an ontology from language f into language t , all labels in language t were removed from the ontology. These eliminated labels were then used as reference standard to evaluate the translation proposals made by the system with respect to standard metrics used in MT research. The evaluation of machine translation is a difficult task as there are many possible valid translations for an input label and the reference translation we use to evaluate the score represents only one possible translation. Thus, we use a collection of widely used evaluation metrics. The list of metrics is as follows:

BLEU Bilingual Evaluation Understudy [9]

METEOR Metric for Evaluation of Translation with Explicit Ordering

[20] NIST The metric as used in the NIST evaluations [21]

PER Position-independent error rate [22]

Source	Language Pair	Translations
Wikipedia	English ↔ German	7,388
	English ↔ Spanish	5,726
	English ↔ Dutch	5,488
IATE	English → Spanish	2,157
	Spanish → English	1,736
	English → German	6,815
	German → English	3,912
	English → Dutch	2,917
	Dutch → English	3,851

Table 4: Number of extra relevant translations found from sources

TER Translation edit rate [23]

WER Word error rate [22]

As we have found in previous studies that BLEU can be unfairly sensitive to short labels, leading to a poorer correlation with human judgement of translations [4], we also introduced a modified version of BLEU, i.e. BLEU-2, which considers only the precision of 1 and 2-grams in evaluating translations. The reason why the BLEU metric is not suited to our task is that it computes the product of the precision for 1,2,3, and 4-grams as an aggregate. However, many of our labels consist of less than 4 words, which would lead to BLEU scores that are zero if the 4-gram precision is zero (as BLEU is a product of each precision score), which is frequently the case if we have a very small sample size for 4-grams. BLEU scores are difficult to interpret but it is generally believed that translations under 0.15 are of too poor quality to be useful in any application and it has reported that the performance of human translators in Wizard of Oz settings is 0.65-0.75 [23].

It is important to note that as PER,TER and WER are error rates, smaller values represent better translation quality.

6.3. Experiments

For generating the translation models from source to target language, we used the statistical translation toolkit Moses [6]. As we aimed to improve the translations only on the surface level, we did not use any additional processing modules of Moses. Word alignments were built with the GIZA++ toolkit [24], where the 5-gram language model was built by SRILM with Kneser-Ney smoothing [25].

For our experiments we used each of these systems developed on the datasets as described above. For the Linguee approach, we built a parallel corpus with around 24,247 aligned sentences for the German GAAP ontology. The number of extra translations generated in domain lexicons is shown in Table 4.

The CL-ESA model used 51,093 articles from a Wikipedia snapshot (October 2012), which were selected on the basis of their length and availability in all of the four considered languages (English, Spanish, German, and Dutch). The articles having less than 100 words were discarded. We tokenized and lower-
 365 cased the remaining articles, removed the stop words, and applied a stemmer before indexing.

The language model was adapted to a set of 452,754 Wikipedia articles in English and Spanish, and 456,496 articles in English and German using the method described in Section 5. These articles were those that were linked to
 370 another article in the other language by means of the ‘articles in other languages’ link and contained at least 100 words in each language. Terms with a frequency under 5 were replaced with an unknown token symbol in the corpus.

6.4. Results for combined system

We present the results for our systems as follows:

375 **Baseline** The baseline system trained with the general purpose Europarl corpus [26]

Wikipedia Lexicon Using the extra terms extracted from Wikipedia as described in section 3.1.1

Linguee + Wikipedia Using extra terminology from Linguee as described in
 380 sections 3.1.2 and 3.1.1

IATE Using extra terminology from IATE as described in section 3.1.3 CLESA

Using the explicit semantic analysis feature as described in section 4.1 LM

Using the language model adaptation procedure as described in section 5

All The combination of the LM adaptation, CLESA features, IATE and either
 385 the Linguee or Wikipedia Lexicon

Table 5 shows the results of the different settings on the financial ontologies, translating from English to Spanish as well as Spanish to English. Table 6 shows the results on the financial ontologies, translating from English to German and German to English. Table 7 show the results of the evaluation when translating
 390 the public service ontologies from English to Dutch and Dutch to English. For each experiment we verify the significance of the improvement of the systems by means of bootstrap resampling [27] and found the improvement between the baseline system and the combined (‘All’) system to be significant at a 99% level. For the financial translation from English to Spanish we see that the com-
 395 bined system outperformed each of the other systems individually, and similarly we see for English and German that nearly all metrics (except for one) show an improvement for the combined system. We further note that the largest single improvement comes from the domain lexicon approach, as discussed below.

For the public service ontologies we see a much less clear result, although in
400 this case there is still a notable improvement for the domain lexicon method. Our
hypothesis for this difference in results is that the public service ontologies contain
rather general and not particularly domain-specific language. As a corollary, this
leads to the hypothesis that our methods are appropriate in the context of
ontologies containing very specific domain terminology. Further, we
405 note that the register of some of the public service texts differ from the training
material in particular in the use of informal forms (e.g., the Dutch pronoun 'je')
and this may have affected the impact of the other methods.

6.5. Discussion

The goal of this evaluation was to compare a baseline SMT system to a SMT 410 system
extended by the different components proposed in this paper, both in isolation and in
combination. According to the results we observe the following:

1. When domain lexicon adaptation is not applied, the results given by the other
techniques in isolation do not differ significantly from the baseline and it is not possible
to conclude which option is best in general.
 - 415 2. When domain lexicon adaptation was applied, the results improved with
respect to the baseline. In this setting, the addition of the other techniques (CLESA,
LM) improves the results even further.
 3. The best results are generally produced by the combination of all the techniques.
- 420 Therefore, we can see that domain lexicon adaptation improves the transla-
tion results most notably, but also serves as an activator of the other techniques. It
should be noted that CLESA and LM Adaptation do not produce new candidate
translations themselves; instead, they aim to optimize the selection of the translation
candidates, which is possible when translation candidates of a better
425 quality are produced by the addition of a domain lexicon into the system.

7. Related Work

Ontology localization consists of two main tasks. The first task involves finding
an appropriate translation for the lexical layer of the ontology, i.e., for all the labels
in the ontology. The second task is the adaptation of the ontology
430 to the – possibly slightly different – conceptualization of the target community that
is supposed to use the localized ontology. In this paper we have been concerned
with the first task only. A previous system concerned with this task was the
LabelTranslator system [28], which was developed as a plug-in for the NEON
project⁷. The LabelTranslator system essentially relied on a rule-based
435 approach as well as many translation candidates collected from external resources
and web services such as Google Translate, EuroWordNet [29] and KMI

⁷<http://www.neon-project.org>

English to Spanish

Method	BLEU	BLEU-2	METEOR	NIST	PER	TER	WER
Baseline	0.114	0.257	0.282	3.644	0.576	0.620	0.676
Linguee + Wikipedia	0.288	0.431	0.444	5.840	0.434	0.441	0.520
IATE	0.113	0.263	0.285	3.663	0.579	0.611	0.670
CLESA	0.113	0.264	0.287	3.683	0.578	0.610	0.670
LM	0.112	0.257	0.279	3.619	0.578	0.620	0.676
All	0.369	0.500	0.496	6.732	0.394	0.378	0.454

Spanish to English

Method	BLEU	BLEU-2	METEOR	NIST	PER	TER	WER
Baseline	0.133	0.302	0.370	3.948	0.642	0.701	0.773
Linguee + Wikipedia	0.414	0.567	0.587	7.139	0.351	0.346	0.463
IATE	0.138	0.310	0.383	4.028	0.628	0.692	0.768
CLESA	0.134	0.303	0.370	3.944	0.642	0.701	0.772
LM	0.131	0.299	0.366	3.904	0.646	0.706	0.777
All	0.438	0.590	0.608	7.428	0.336	0.327	0.440

Table 5: Results for translating ontology labels from English to Spanish and Spanish to English on the financial ontologies using different settings.

English to German

Method	BLEU	BLEU-2	METEOR	NIST	PER	TER	WER
Baseline	0.064	0.172	0.202	2.596	0.813	0.842	0.861
Linguee + Wikipedia	0.186	0.312	0.337	4.451	0.656	0.691	0.716
IATE	0.060	0.159	0.193	2.466	0.828	0.855	0.871
CLESA	0.062	0.167	0.200	2.533	0.814	0.841	0.860
LM	0.064	0.172	0.202	2.596	0.812	0.841	0.860
All	0.208	0.334	0.360	4.694	0.632	0.660	0.683

German to English

Method	BLEU	BLEU-2	METEOR	NIST	PER	TER	WER
Baseline	0.085	0.218	0.274	3.029	0.797	0.814	0.870
Linguee + Wikipedia	0.271	0.420	0.447	5.601	0.553	0.520	0.634
IATE	0.086	0.217	0.269	3.015	0.794	0.811	0.866
CLESA	0.086	0.217	0.272	3.022	0.796	0.812	0.868
LM	0.081	0.218	0.273	3.019	0.798	0.814	0.870
All	0.279	0.428	0.444	5.782	0.542	0.505	0.617

Table 6: Results for translating ontology labels from English to German and German to English on the financial ontologies using different settings.

English to Dutch

Method	BLEU	BLEU-2	METEOR	NIST	PER	TER	WER
Baseline	0.176	0.202	0.152	2.011	0.849	0.745	0.863
Wikipedia Lexicon	0.252	0.259	0.183	2.467	0.778	0.673	0.795
IATE	0.172	0.197	0.146	1.996	0.845	0.748	0.864
CLESA	0.142	0.176	0.143	1.969	0.857	0.754	0.876
LM	0.175	0.200	0.152	2.009	0.849	0.747	0.863
All	0.167	0.210	0.151	2.216	0.835	0.720	0.851

Dutch to English

Method	BLEU	BLEU-2	METEOR	NIST	PER	TER	WER
Baseline	0.172	0.238	0.200	2.511	0.774	0.670	0.789
Wikipedia Lexicon	0.211	0.269	0.224	2.814	0.720	0.619	0.734
IATE	0.178	0.238	0.198	2.540	0.778	0.674	0.797
CLESA	0.179	0.236	0.196	2.481	0.782	0.679	0.796
LM	0.172	0.238	0.200	2.509	0.774	0.671	0.789
All	0.224	0.275	0.223	2.879	0.719	0.622	0.734

Table 7: Results for translating ontology labels from English to Dutch and Dutch to English on the public services ontologies using different settings.

Watson [30] as well as techniques for ranking these translations given the ontological context. An important bottleneck is the limited availability of online web translation systems and online dictionaries. LabelTranslator used basic lexical template rules to attempt to tackle this sparsity issue by means of limited transfer grammars, which does not generalize well.

For ontology localization, a similar task is the one of finding an alignment between existing ontologies which have different conceptualizations of the same domain but in different languages. These tasks frequently build on a label translation system, and thus convert the task of finding a cross-lingual alignment to that of finding a monolingual alignment among translated labels [31, 32]. In a similar direction, Carpuat et al. [33] derived cross-lingual alignments between English and Chinese WordNet by means of a distributional similarity approach.

There has also been much research in the process of finding translingual semantic representations, starting with methods that merge parallel corpora to obtain a translingual representation by means of latent topic modelling methods such as Latent Semantic Analysis [34] and Latent Dirichlet Allocation [35]. These methods can be used to estimate the latent similarity between ontology labels in different languages, but until recently such approaches did not yield results that outperformed direct translation [36] in comparable tasks. In particular, recent methods such as Orientated Principle Component Analysis [36], Kernel Canonical Correlation Analysis [37, CCA] and Orthonormal Explicit Topic Analysis [38, ONETA] calculate a translingual representation by means of estimating the correlation between term frequencies in a document-aligned corpus and have been shown to outperform dictionary-based machine translation [39].

There have been several systems developed for the adaptation of machine translation systems to different domains. For example, Koehn and Schroeder [40] showed that the quality of a machine translation system can be improved by
 465 interpolating a small amount of in-domain knowledge into either the language model or the translation model. Another approach consists in using in-domain text along with a cross-lingual similarity method to mine translations that are specific to the domain, for example by the use of CCA [41]. A similar approach is to use a small amount of (unaligned) in-domain text to sample an in-domain
 470 section of a larger parallel corpus [42]. Furthermore, there are many approaches that are designed to adapt language models to domains. Bellegarda [43] characterizes these into three main classes: firstly, interpolation models, which attempt to combine a small in-domain language model with a larger general domain model (e.g., [40]). Secondly, Bellegarda describes constraint-based models
 475 which combine a general domain model with an in-domain model while maximizing some criteria, generally Minimum Discriminative Information [44]. Finally, Bellegarda describes models, which extend existing topic models, to give probabilities for n -grams not just bag of words [45]. Nevertheless, none of such methods explicitly exploit the ontological context and are specifically designed
 480 to deal with short labels.

We note that since the submission of this article one of the authors has continued to develop some of the methods presented in this paper into the OTTO translation system that can be used online ⁸ [46, 47].

8. Conclusion

485 We have tackled the problem of ontology localization/translation and presented a framework for domain adaptation which factors in the ontological context to provide appropriate translations of labels in an ontology. The domain adaptation framework has been implemented on top of an existing state-of-the-art and off-the-shelf machine translation system. We have in particular
 490 presented three techniques for domain adaptation to a given ontology: i) extraction of a bilingual dictionary from external resources such as Wikipedia, from existing parallel corpora as well as from third-party terminologies, ii) exploiting a semantic similarity measure to rerank translation candidates, thus supporting disambiguation, and iii) tuning of a monolingual language model to
 495 the ontology. We have presented experiments on five ontologies showing the impact of each of these methods. As one interesting result we have shown that the methods used in combination indeed improve the quality of translations. However, the above mentioned techniques ii) and iii) only yield an impact if method i) is used as well. The reason for this is essentially that methods ii)
 500 and iii) are ranking/scoring techniques that can perform domain-specific disambiguation, but can not introduce new translations themselves. Method i) in contrast introduces new domain-specific translation candidates that can then

⁸<http://server1.nlp.insight-centre.org/otto/>

be assigned a score to which methods ii) and iii) contribute. Overall, we have presented a new methodology for tuning an off-the-shelf statistical translation system to the task of translating the labels of a given ontology into another language. All software is available as an extension to the Moses system so that our methodology is of high practical value as it can be used by any third party.

Our experiments have shown in particular that the approach of building a new, domain-specific corpus showed a large impact on the translation quality. Further, our approach has shown that collaboratively created resources such as Wikipedia and DBpedia can be successfully exploited to tune an SMT system and provide higher quality translations in ontology localization tasks. In addition to Wikipedia article titles with their multilingual equivalents, Wikipedia holds much more information in the articles themselves. Further work should investigate how to exploit such non-parallel resources to improve the performance of SMT systems.

Acknowledgements

This research was supported in by funding from the Monnet project under European Union FP7 program under grant number 248458, the CITEC excellence initiative funded by the DFG (Deutsche Forschungsgemeinschaft), the Science Foundation Ireland under Grant Number SFI/12/RC/2289 (Insight).

References

- [1] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, J. McCrae, Challenges for the multilingual Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web 11 (2012) 63–71.
- [2] P. Buitelaar, K.-S. Choi, P. Cimiano, E. H. Hovy, The Multilingual Semantic Web (Dagstuhl Seminar 12362), Dagstuhl Reports 2 (9) (2013) 15–94. doi:<http://dx.doi.org/10.4230/DagRep.2.9.15>.
- [3] A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia, G. Aguado-de Cea, Guidelines for multilingual linked data, in: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, WIMS’13, New York, NY, USA, 2013, pp. 14–25. doi:10.1145/2479787.2479867.
- [4] J. McCrae, E. Montiel-Ponsoda, G. Aguado de Cea, M. J. Espinoza Mejía, P. Cimiano, Combining statistical and semantic approaches to the translation of ontologies and taxonomies, in: Proceedings of 7th Workshop on Syntax, Structure and Semantics in Statistical Translation, 2011, pp. 116–125.
- [5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: A nucleus for a web of open data, in: The Semantic Web, Springer, 2007, pp. 722–735.

- [6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al., Moses: Open source toolkit for statistical machine translation, in: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, 2007, pp. 177–180.
- [7] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis, in: Proceedings of the 20th international joint conference on artificial intelligence, Vol. 6, 2007, p. 12.
- [8] P. Sorg, P. Cimiano, Cross-lingual information retrieval with explicit semantic analysis, in: Working Notes for the CLEF 2008 Workshop, 2008.
- [9] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318.
- [10] P. Koehn, Statistical machine translation, Cambridge University Press, 2010.
- [11] U. Hermjakob, K. Knight, H. DauméIII, Name translation in statistical machine translation-learning when to transliterate., in: ACL, 2008, pp. 389–397.
- [12] P. Koehn, F. J. Och, D. Marcu, Statistical phrase-based translation, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Association for Computational Linguistics, 2003, pp. 48–54.
- [13] A. Miles, S. Bechhofer, SKOS Simple Knowledge Organization System, W3c recommendation, World Wide Web Consortium (2009).
- [14] T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse processes 25 (2-3) (1998) 259–284.
- [15] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
- [16] M. Carpuat, D. Wu, Evaluating the word sense disambiguation performance of statistical machine translation, in: Proceedings of the second international joint conference on natural language processing (IJCNLP), 2005, pp. 122–127.
- [17] M. Carpuat, D. Wu, Improving statistical machine translation using word sense disambiguation, in: In The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, pp. 61–72.

- 580 [18] S. F. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, in: Proceedings of the 34th annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1996, pp. 310–318.
- [19] Phillip Engel and Walter Hamscher and Geoffrey Shuetrim and David
585 von Kannon and Hugh Wallis, Extensible Business Reporting Language (XBRL) 2.1, Tech. rep., XBRL International (2003).
URL <http://www.xbrl.org/Specification/XBRL-2.1/REC-2003-12-31/XBRL-2.1-REC-2003-12-31+corrected-errata-2013-02-20.html>
- 590 [20] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.
- [21] G. Doddington, Automatic evaluation of machine translation quality using
595 n-gram co-occurrence statistics, in: Proceedings of the Second International Conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc., 2002, pp. 138–145.
- [22] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, H. Sawaf, Accelerated DP based search for statistical translation, in: European Conf. on Speech Commu-
600 nication and Technology, 1997, pp. 2667–2670.
- [23] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the Association for Machine Translation in the Americas, 2006, pp. 223–231.
- 605 [24] F. J. Och, H. Ney, A systematic comparison of various statistical alignment models, Computational Linguistics 29 (1) (2003) 19–51.
- [25] A. Stolcke, SRILM-an extensible language modeling toolkit, in: Proceedings International Conference on Spoken Language Processing, 2002, pp. 257–286.
- 610 [26] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: MT summit, Vol. 5, 2005, pp. 79–86.
- [27] P. Koehn, Statistical significance tests for machine translation evaluation., in: EMNLP, 2004, pp. 388–395.
- [28] M. Espinoza, A. Gómez-Pérez, E. Mena, Labeltranslator-a tool to auto-
615 matically localize an ontology, in: The Semantic Web: Research and Applications, Springer, 2008, pp. 792–796.
- [29] P. Vossen, EuroWordNet: a multilingual database with lexical semantic networks, Kluwer Academic Boston, 1998.

- [30] M. d'Aquin, L. Gridinoc, S. Angeletou, M. Sabou, E. Motta, Watson: A gateway for next generation semantic web applications, in: Proceedings of the 6th International Semantic Web Conference, 2007.
- [31] B. Fu, R. Brennan, D. O'Sullivan, Cross-lingual ontology mapping—an investigation of the impact of machine translation, in: The Semantic Web, Springer, 2009, pp. 1–15.
- [32] D. Spohr, L. Hollink, P. Cimiano, A machine learning approach to multilingual and cross-lingual ontology matching, in: The Semantic Web—ISWC 2011, Springer, 2011, pp. 665–680.
- [33] M. Carpuat, G. Ngai, P. Fung, K. Church, Creating a bilingual ontology: a corpus-based approach for aligning wordnet and HowNet, in: Proceedings of the 1st Global WordNet Conference, 2002.
- [34] S. T. Dumais, T. A. Letsche, M. L. Littman, T. K. Landauer, Automatic cross-language retrieval using latent semantic indexing, in: AAAI spring symposium on cross-language text and speech retrieval, Vol. 15, 1997, p. 21.
- [35] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, A. McCallum, Polylingual topic models, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, Association for Computational Linguistics, 2009, pp. 880–889.
- [36] J. C. Platt, K. Toutanova, W.-T. Yih, Translingual document representations from discriminative projections, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010, pp. 251–261.
- [37] A. Vinokourov, N. Cristianini, J. S. Shawe-taylor, Inferring a semantic representation of text via cross-language correlation analysis, in: Advances in Neural Information Processing Systems, 2002, pp. 1473–1480.
- [38] J. McCrae, P. Cimiano, R. Klinger, Orthonormal explicit topic analysis for cross-lingual document matching, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1732–1740.
- [39] J. Jagarlamudi, R. Udupa, H. DauméIII, A. Bhole, Improving bilingual projections via sparse covariance matrices, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 930–940.
- [40] P. Koehn, J. Schroeder, Experiments in domain adaptation for statistical machine translation, in: Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2007, pp. 224–227.

- 660 [41] H. DauméIII, J. Jagarlamudi, Domain adaptation for machine translation by mining unseen words., in: Proceedings of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 407–412.
- [42] A. Axelrod, X. He, J. Gao, Domain adaptation via pseudo in-domain data selection, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 355–362.
- 665 [43] J. R. Bellegarda, Statistical language model adaptation: review and perspectives, *Speech communication* 42 (1) (2004) 93–108.
- [44] M. Federico, Efficient language model adaptation through MDI estimation., in: Proceedings of the Eurospeech Conference, 1999, pp. 1583–1586.
- [45] H. M. Wallach, Topic modeling: beyond bag-of-words, in: Proceedings of the 23rd International Conference on Machine learning, 2006, pp. 977–984.
- 670 [46] M. Arcan, M. Turchi, P. Buitelaar, Knowledge portability with semantic expansion of ontology labels, in: The 53rd Annual Meeting of the Association for Computational Linguistics, 2015, pp. 708–718.
- [47] M. Arcan, K. Asooja, H. Ziad, P. Buitelaar, OTTO – Ontology Translation System, in: ISWC 2015 Posters & Demonstrations Track, 2015

