Knowledge Portability with Semantic Expansion of Ontology Labels

Mihael Arcan[1] **Marco Turchi**[2] **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
turchi@fbk.eu

Abstract

Our research focuses on the multilingual enhancement of ontologies that, often represented only in English, need to be translated in different languages to enable knowledge access across languages. Ontology translation is a rather different task then the classic document translation, because ontologies contain highly specific vocabulary and they lack contextual information. For these reasons, to improve automatic ontology translations, we first focus on identifying relevant unambiguous and domain-specific sentences from a large set of generic parallel corpora. Then, we leverage Linked Open Data resources, such as DBPedia, to isolate ontology-specific bilingual lexical knowledge. In both cases, we take advantage of the semantic information of the labels to select relevant bilingual data with the aim of building an ontology-specific statistical machine translation system. We evaluate our approach on the translation of a medical ontology, translating from English into German. Our experiment shows a significant improvement of around 3 BLEU points compared to a generic as well as a domain-specific translation approach.

1 Introduction

Currently, most of the semantically structured data, i.e. ontologies or taxonomies, has labels expressed in English only.[1] On the one hand, the increasing amount of ontologies offers an excellent opportunity to link this knowledge together (Gómez-Pérez et al., 2013). On the other hand, non-English users may encounter difficulties when using the ontological knowledge represented only in English. Furthermore, applications in information retrieval, question answering or knowledge management, that use monolingual ontologies are therefore limited to the language in which the ontology labels are stored. To make the ontological knowledge language-independent and accessible beyond language borders, these monolingual resources need to be transformed into multilingual knowledge bases. This multilingual enhancement can enable queries on documents beyond English, e.g. for cross-lingual business intelligence in the financial domain (O'Riain et al., 2013), providing information related to an ontology label, e.g. *other intangible assets*,[2] in Spanish, German or Italian. The main challenge involved in building multilingual knowledge bases is, however, to bridge the gap between language-specific information and the language-independent semantic content of ontologies or taxonomies (Gracia et al., 2012).

Since manual multilingual enhancement of ontologies is a very time consuming and expensive process, we engage an ontology-specific statistical machine translation (SMT) system to automatically translate the ontology labels. Due to the fact that ontology labels are usually highly domain-specific and stored only in knowledge representations (Chandrasekaran et al., 1999), the labels appear infrequent in parallel corpora, which are needed to build a domain-specific translation system with accurate translation candidates. Additionally, ambiguous labels built out of only a few words do often not express enough semantic or contextual information to guide the SMT system to translate a label into the targeted domain. This can be observed by domain-unadapted SMT systems, e.g. Google Translate, where ambiguous expressions, such as *vessel* stored in an medical ontology, are often translated into a generic do-

[1] Based on (Gracia et al., 2012), around 80% of ontology labels indexed in Watson are English.

[2] ontology label stored in FINREP - FINancial REPorting

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 708–718, Beijing, China, July 26-31, 2015.

main as *Schiff*[3] in German (meaning *ship* or *boat*), but not into the targeted medical domain as *Gefäß*. Since ontologies may change over time, keeping up with these changes can be challenging for a human translator. Having in place an SMT system adapted to an ontology can therefore be very beneficial.

In this work, we propose an approach to select the most relevant (parallel) sentences from a pool of generic sentences based on the lexical and semantic overlap with the ontology labels. The goal is to identify sentences that are domain-specific in respect of the target domain and contain as much as possible relevant words that can allow the SMT system to learn the translations of the monolingual ontology labels. For instance, with the sentence selection we aim to retain only parallel sentences where the English word *injection* is translated into the German language as *Impfung* in the medical domain, but not into *Eindüsung*, belonging to the technical domain. This selection process aims to reduce the semantic noise in the translation process, since we try to avoid learning translation candidates that do not belong to the targeted domain. Nonetheless, some of the domain-specific ontology labels may not be automatically translatable with SMT, due to the fact that the bilingual information is missing and cannot be learned from the parallel sentences. Therefore we use the information contained in the DBpedia knowledge base (Lehmann et al., 2015) to improve the translation of expressions which are not known to the SMT system. We tested our approach on the medical domain translating from English to German, showing improvements of around 3 BLEU points compared to a generic as well as a domain-specific translation model.

The remainder of this paper is organized as follows: Section 2 gives an overview of the related work done in the field of ontology translation within SMT. In Section 3, we present the methodology of parallel data selection and terminology identification to improve ontology label translation. Furthermore we show different methods of embedding domain-specific knowledge into SMT. In Experimental Setting, Section 4, we describe the ontology to be translated along the training data needed for SMT. Moreover we introduce existing approaches and give a description of metrics for automatic translation evaluation. Section 5 presents the automatic and manual evaluation of the translated labels. Finally, conclusions and future work are shown in Section 6.

[3]Translation performed on 25.02.2015

2 Related Work

The task of ontology translation involves the finding of an appropriate translation for the lexical layer, i.e. labels, of the ontology. Most of the previous work tackled this problem by accessing multilingual lexical resources, e.g. EuroWordNet or IATE (Declerck et al., 2006; Cimiano et al., 2010). Their work focuses on the identification of the lexical overlap between the ontology and the multilingual resource. Since the replacement of the source and target vocabulary guarantees a high precision but a low recall, external translation services, e.g. BabelFish, SDL FreeTranslation tool or Google Translate, were used to overcome this issue (Fu et al., 2009; Espinoza et al., 2009). Additionally, ontology label disambiguation was performed by (Espinoza et al., 2009) and (McCrae et al., 2011), where the structure of the ontology along with existing multilingual ontologies 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 using a small, but ontology-specific SMT system. We learned that using external SMT services often results in wrong translations of labels, because the external SMT services are not able to adapt to the specificity of the ontology. Avoiding existing multilingual resources, which enables a simple replacement of source and target labels, showed the possibility of improving label translations without manually generated lexical resources, since not every ontology may benefit of current multilingual resources.

Due to the specificity of the labels, previous research (Wu et al., 2008; Haddow and Koehn, 2012) showed that generic SMT systems, which merge all accessible data together, cannot be used to translate domain-specific vocabulary. To avoid unsatisfactory translations of specific vocabulary we have to provide the SMT system domain-specific bilingual knowledge, from where it can learn specific translation candidates. (Eck et al., 2004) used for the language model adaptation within SMT the information retrieval technique *tf-idf*. Similarly, (Hildebrand et al., 2005) and (Lü et al., 2007) utilized this approach to select

relevant sentences from available parallel text to adapt translation models. The results confirmed that large amounts of generic training data cannot compensate for the requirement of domain-specific training sentences. Another approach is taken by (Moore and Lewis, 2010), where, based on source and target language models, the authors calculated the difference of the cross-entropy values for a given sentence. (Axelrod et al., 2011) extend this work using the bilingual difference of cross-entropy on in-domain and out-of-domain language models for training sentence selection for SMT. (Wuebker et al., 2014) reused the cross-entropy approach and applied it to the translation of video lectures. (Kirchhoff and Bilmes, 2014) introduce submodular optimization using complex features for parallel sentence selection. In their experiments they use the source and target side of the text to be translated, and show significant improvements over the widely used cross-entropy method. A different approach for sentence selection is shown in (Cuong and Sima'an, 2014), where the authors propose a latent domain translation model to distinguish between hidden in- and out-of-domain data. (Gascó et al., 2012) and (Bicici and Yuret, 2011) sub-sample sentence pairs whose source has most overlap with the evaluation dataset. Different from these approaches, we do not embed any specific in-domain knowledge to the generic corpus, from which sentence selection is performed. Furthermore, none of these methods explicitly exploit the ontological hierarchy for label disambiguation and are not specifically designed to deal with the characteristics of ontology labels.

As a lexical resource, Wikipedia with its rich semantic knowledge was used as a resource for bilingual term identification in the context of SMT. (Tyers and Pieanaar, 2008) extracts bilingual dictionary entries from Wikipedia to support the machine translation system. Based on exact string matching they query Wikipedia with a list of around 10,000 noun lemmas to generate the bilingual dictionary. Besides the interwiki link system, (Erdmann et al., 2009) enhance their bilingual dictionary by using redirection page titles and anchor text within Wikipedia. To cast the problem of ambiguous Wikipedia titles, (Niehues and Waibel, 2011; Arcan et al., 2014a) use the information of Wikipedia categories and the text of the articles to provide the SMT system domain-specific bilingual knowledge. This research showed that using the lexical information stored in this knowledge base improves the translation of highly domain-specific vocabulary. However, we do not rely on category annotations of Wikipedia articles, but perform domain-specific dictionary generation based on the overlap between related words from the ontology label and the abstract of a Wikipedia article.

3 Methodology

We propose an approach that uses the ontology labels to be translated to select the most relevant parallel sentences from a generic parallel corpus. Since ontology labels tend to be short (McCrae et al., 2011), we expand the label representation with its semantically related words. This expansion enables a larger semantic overlap between a label and the (parallel) sentences, which gives us more information to distinguish between related and unrelated sentences. Our approach reduces the ambiguity of expressions in the selected parallel sentences, which consequently gives more preference to translation candidates of the targeted domain. Furthermore, we access the DBpedia knowledge base to identify bilingual terminology belonging to the domain of the ontology. Once the domain-specific parallel sentences and lexical knowledge is available, we use different techniques to embed this knowledge into the SMT system. These methods are detailed in the following subsections.

3.1 Domain-Specific Parallel Sentence Selection

In order to generate the best translation system we select only sentences from the generic parallel corpus which are most relevant to the labels to be translated. The first criteria for relevance was the *n-gram overlap* between a label and a source sentence coming from the generic corpus. Therefore we calculate the cosine similarity between the n-grams extracted from a label and the n-grams of each source sentence in the generic corpus. The similarity between the label and the sentence is defined as the cosine of the angle between the two vectors. The calculated similarity score allows us to distinguish between more and less relevant sentences.

Due to the specificity of ontology labels, the *n-gram overlap* approach is not able to select useful sentences in the presence of short labels. For

this reason, we improve it by extending the semantic information of labels using a technique for computing vector representations of words. The technique is based on a neural network that analyses the textual data provided as input and provides as output a list of semantically related words (Mikolov et al., 2013). Each input string is vectorized using the surrounding context and compared to other vectorized sets of words (from the training data) in a multi-dimensional vector space. For obtaining the vector representations we used a distributional semantic model trained on the Wikipedia articles,[4] containing more than 3 billion words. Word relatedness is measured through the cosine similarity between two word vectors. A score of 1 would represent a perfect word similarity; e.g. *cholera* equals *cholera*, while the medical expression *medicine* has a cosine distance of 0.678 to *cholera*. Since words, which occur in similar contexts tend to have similar meanings (Harris, 1954), this approach enables to group related words together. The output of this technique is the analysed label with a vector attached to it, e.g. for the medical label *cholera* it provides related words with its relatedness value, e.g. *typhus* (0.869), *smallpox* (0.849), *epidemic* (0.834), *dysentery* (0.808) ... In our experiments, this method is implemented by the use of Word2Vec.[5]

To additionally disambiguate short labels, the related words of the current label are combined with the related words of its direct parent in the ontology. The usage of the ontology hierarchy allows us to take advantage of the specific vocabulary of the related words in the computation of the cosine similarity. 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 passed to the cosine similarity. The most similar source sentence and the label should share the largest number of related words (largest cosine similarity).

3.2 Bilingual Terminology Identification

The automatic translation of domain-specific vocabulary can be a hard task for a generic SMT system, if the bilingual knowledge is not present in the parallel dataset. To complement the previous approaches we access DBpedia[6] as a multilingual lexical resource.

We engage the idea of (Arcan et al., 2012) where the authors provide to the SMT system unambiguous terminology identified in Wikipedia to improve the translations of labels in the financial domain. To disambiguate Wikipedia entries with translations into different domains, they query the repository for analysing the n-gram overlap between the financial labels and the Wikipedia entries and store the frequency of categories which are associated with the matched entry. In a final step they extract only bilingual Wikipedia entries, which are associated with the most frequent Wikipedia categories identified in the previous step.

Since the Wikipedia entries are often associated only with a few categories, this limited vocabulary may give only a small contribution for this disambiguation of different meanings or topics of the same Wikipedia entry. For this reason, we use for each Wikipedia entry the extended abstract, which contains more information about the entry compared to the previous approach. For ambiguous Wikipedia entries, which overlap with a medical label, we therefore calculate the cosine similarity between the related words associated with the label and the lexical information of the Wikipedia abstract. Among different ambiguous entries, the cosine similarity gives more weight to the Wikipedia entry, which is closer to our preferred domain. Finally, if the Wikipedia entry has an equivalent in the target language, i.e. German, we use the bilingual information for the lexical enhancement of the SMT system.

3.3 Integration of Domain-Specific Knowledge into SMT

After the identification of domain-specific bilingual knowledge, it has to be integrated into the workflow of the SMT system. The injection of new obtained knowledge can be performed by retraining the domain-specific knowledge with the generic parallel corpus (Langlais, 2002; Ren et al., 2009; Haddow and Koehn, 2012) or by adding new entries directly to the translation system (Pinnis et al., 2012; Bouamor et al., 2012). These methods have the drawback that the bilingual domain specificity may get lost due to the usually larger generic parallel corpora. Giving more priority to domain-specific translations than generic ones, we focus on two techniques, i.e. the Fill-Up model (Bisazza et al., 2011) and the Cache-Based

[4] Wikipedia dump id enwiki-20141106

[5] https://code.google.com/p/word2vec/

[6] http://wiki.dbpedia.org/Downloads2014

Model (Bertoldi et al., 2013) approach.

The Fill-Up model has been developed to address a common scenario where a large generic background model exists, and only a small quantity of domain-specific data can be used to build a translation model. Its goal is to leverage the large coverage of the background model, while preserving the domain-specific knowledge coming from the domain-specific data. For this purpose the generic and the domain-specific translation models are merged. For those translation candidates that appear in both models, only one instance is reported in the Fill-Up model with the largest probabilities according to the translation models. To keep track of a translation candidate's provenance, a binary feature is added that gives preference to a translation candidate if it comes from the domain-specific translation model. We engage the idea of the Fill-Up model to combine the domain-specific parallel knowledge from the selected sentences with the generic (1.9M) parallel corpus.

Furthermore, for embedding bilingual lexical knowledge into the SMT system, we engage the idea of cache-based translation and language models (Bertoldi et al., 2013). The main idea behind these models is to combine a large static global model with a small, but dynamic local model. This approach has already shown its potential of injecting domain-specific knowledge into a generic SMT system (Arcan et al., 2014b). For our experiments we inject the bilingual lexical knowledge identified in DBpedia and IATE into the cache-based models. The cache-based model relies on a local translation model (CBTM) and language model (CBLM). The first is implemented as an additional table in the translation model providing one score. All entries are associated with an 'age' (initially set to 1), corresponding to the time when they were actually inserted. Each new insertion causes an ageing of the existing translation candidates and hence their re-scoring; in case of re-insertion of a phrase pair, the old value is set to the initial value. Similarly to the CBTM, the local language model is built to give preference to the provided target expressions. Each entry stored in CBLM is associated with a decaying function of the age of insertion into the model. Both models are used as additional features of the log-linear model in the SMT system.

4 Experimental Setting

In this Section, we give an overview on the dataset and the translation toolkit used in our experiment. Furthermore, we describe the existing approaches and give insights into the SMT evaluation techniques, considering the translation direction from English to German.

Evaluation Dataset For our experiments we used the International Classification of Diseases (ICD) ontology as the gold standard,[7] whereby the considered translation direction is from English to German. The ICD ontology, translated into 43 languages, is used to monitor diseases and to report the general health situation of the population in a country. This stored information also provides an overview of the national mortality rate and appearance of diseases of WHO member countries.

For our experiment we used 2000 English labels from the ICD-10 dataset, which were aligned to their German equivalents (Table 1). To identify the best set of sentences we experiment with different values of τ, which is the percentage of all the sentences that are considered relevant (domain-specific) by the sentence extraction approach. The value that allows the SMT system to achieve the best performance on the *development dataset 1* is used on the *evaluation set*, which is used for the translation evaluation of ontology labels reported in this paper. The parameters within the SMT system are optimized on the *development dataset 2*.

Statistical Machine Translation and Training Dataset For our translation task, we use the statistical translation toolkit Moses (Koehn et al., 2007), where the word alignments were built with the GIZA++ toolkit (Och and Ney, 2003). The SRILM toolkit (Stolcke, 2002) was used to build the 5-gram language model.

For a broader domain coverage of the generic training dataset necessary for the SMT system, we merged parts of JRC-Acquis 3.0[8] (Steinberger et al., 2006), Europarl v7[9] (Koehn, 2005) and OpenSubtitles2013[10] (Tiedemann, 2012), obtaining a training corpus of 1.9M sentences, con-

[7] http://www.who.int/classifications/icd/en/

[8] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis

[9] http://www.statmt.org/europarl/

[10] http://opus.lingfil.uu.se/OpenSubtitles2013.php

		English	German
Generic Dataset (out-domain)	Sentences	1.9M	
	Running Words	39.8M	37.1M
	Vocabulary	195,912	446,068
EMEA Dataset (domain-specific)	Sentences	1.1M	
	Running Words	13.8M	12.7M
	Vocabulary	58,935	115,754
Development Dataset 1	Labels	500	
	Running Words	3,025	2,908
	Vocabulary	889	951
Development Dataset 2	Labels	500	
	Running Words	3,003	3,020
	Vocabulary	938	1,027
Evaluation Dataset	Labels	1,000	
	Running Words	5,677	5,514
	Vocabulary	1,255	1,489

Table 1: Statistics for the bilingual training, development and evaluation datasets. ('Vocabulary' denotes the number of unique words in the dataset)

taining around 38M running words (Table 1).[11] The generic SMT system, trained on the concatenated 1.9 sentences, is used as a baseline, which we compare against the domain-specific models generated with different sentence selection methods. Furthermore we use the generic SMT system in combination with the smaller domain-specific models to evaluate different approaches when combining generic and domain-specific data together.

We additionally compare our results to an SMT system built on an existing domain-specific parallel dataset, i.e. EMEA[12] (Tiedemann, 2009), which holds specific medical parallel data extracted from the European Medicines Agency documents and websites.

Comparison to Existing Approaches We compare our approach on knowledge expansion for sentence selection with similar methods that distinguish between more important sentences and less important ones. First, we sort 1.9M sentences from the generic corpus based on the *perplexity* of the ontology vocabulary. The perplexity score gives a notion of how well the probability model based on the ontology vocabulary predicts a sample, which is in our case each sentence in the generic corpus.

Second, we use the method shown in (Hildebrand et al., 2005), where the authors use a method based on *tf-idf*[13] to select the most relevant sentences. This widely-used method in information retrieval tells us how important a word is to a document, whereby each sentence from the generic corpus is treated as a document.

Finally, we compare our approach with the *infrequent n-gram recovery* method, described in (Gascó et al., 2012). Their technique consists of selection of relevant sentences from the generic corpus, which contain infrequent n-grams based on their test data. They consider an n-gram as infrequent if it appears in the generic corpus less times than an infrequent threshold *t*.

Furthermore we enrich and evaluate our proposed ontology-specific SMT system with the lexical information coming from the terminological database IATE[14] (Inter-Active Terminology for Europe). IATE is the institutional terminology database of the EU and is used for the collection, dissemination and shared management of specific terminology and contains approximately 1.4 million multilingual entries.

Evaluation Metrics The automatic translation evaluation is based on the correspondence between the SMT output and reference translation (gold standard). For the automatic evaluation we used the BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) algorithms.[15]

BLEU (Bilingual Evaluation Understudy) is calculated for individual translated segments (n-grams) by comparing them with a dataset of reference translations. Considering the shortness of the labels, we report scores based on the bi-gram overlap (BLEU-2) and the standard four-gram overlap (BLEU-4). Those scores, between 0 and 100 (perfect translation), are then averaged over the whole *evaluation dataset* to reach an estimate of the translation's overall quality.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. Along with standard exact word (or phrase) matching it has additional features, i.e. stemming, paraphrasing and synonymy matching. Differently to BLEU, the metric produces good correlation with human judgement at the sentence or segment level.

[11] For reproducibility and future evaluation we take the first one-third part of each corpus.

[12] `http://opus.lingfil.uu.se/EMEA.php`

[13] *tf-idf* – term frequency-inverse document frequency

[14] `http://iate.europa.eu/downloadTbx.do`

[15] METEOR configuration: exact, stem, paraphrase

The approximate randomization approach in MultEval (Clark et al., 2011) is used to test whether differences among system performances are statistically significant with a p-value < 0.05.

5 Evaluation of Ontology Labels

In this Section, we report the translation quality of ontology labels based on translation systems learned from different sentence selection methods. Additionally, we perform experiments training an SMT system on the combination of in- and out-domain knowledge. The final approach enhances a domain-specific translation system with lexical knowledge identified in IATE or DBpedia.

5.1 Automatic Translation Evaluation

We report the automatic evaluation based on BLEU and METEOR for the sentence selection techniques, the combination of in- and out-domain data and the lexical enhancement of SMT.

Sentence Selection Techniques As a first evaluation, we automatically compare the quality of the ICD labels translated with different SMT systems trained on specific sentences by the aforementioned selection techniques (Table 2). Due to the in-domain bilingual knowledge, the translation system trained using the EMEA dataset performs slightly better compared to the large generic baseline system. Among the different sentence selection approaches, the *infrequent n-gram recovery* method (infreq. in Table 2) outperforms the baselines and all the other techniques. This is due to the very strict criteria of selecting relevant sentences that allows the *infrequent n-gram recovery* method to identify a limited number (20,000) of highly ontology-specific bilingual sentences. The *related words* and the *n-gram overlap* models perform slightly better than the baseline, with a usage of 81,000 and 59,000 relevant sentences, and perform similarly to the in-domain EMEA translation system.

Further translation quality improvement is possible, if sentence selection methods are combined together (last four rows in Table 2). The cosine similarities of the methods are combined together, whereby new thresholds τ are computed on the *development dataset 1* and applied on the ICD *evaluation dataset*. Each combined method showed improvement compared to the stand-alone method. The best overall performance is obtained when combining the *n-gram overlap*, the semantic *related words* and *infrequent n-gram recovery* methods. With this combination, we reduce the amount of parallel sentences by 98% compared to the generic corpus and significantly outperform the baseline by 2.3 BLEU score points. These two factors confirm the capability of the combined approach of selecting only few ontology-specific bilingual sentences (30,000) that allows the SMT system to identify the correct translations in the target ontology domain. This is due to the fact that the three combined methods are quite complementary. In fact, the *n-gram overlap* method selects a relatively large amount of bilingual sentences with few words in common with the label, the *related words* approach identifies bilingual sentences in the ontology target domain, and the infrequent n-gram recovery technique selects few bilingual sentences with only specific n-grams in common with the labels balancing the effect of the n-gram overlap method.

Dataset Type	Size	BLEU-2	BLEU-4	METEOR
Generic dataset	1.9M	17.2	6.6	24.7
EMEA dataset	1.1M	18.5	7.0	25.8
(1) perplexity	89K	17.5	6.8	24.8
(2) tf-idf	21K	12,6	4.9	18,7
(3) infreq.	20K	19.1	8.1	25.3
(4) related w.	81K	18.9	7.0	25.8
(5) n-gram	59K	17.7	7.1	23.3
(5) $\wedge$ (3)	22K	18.9	8.2*	25.1
(5) $\wedge$ (4)	24K	17.3	7.3	23.9
(3) $\wedge$ (4)	24K	18.4	8.4*	25.5*
(5) $\wedge$ (4) $\wedge$ (3)	30K	**20.1**	**8.9***	**27.2***

Table 2: Automatic translation evaluation on the evaluation dataset of the ICD ontology (Size = amount of selected sentences from the generic parallel corpus. bold results = best performance; *statistically significant compared to baseline)

Combining In- and Out-Domain Data Considering the relatively small amount of parallel data extracted with the sentence selecting methods for the SMT community, we evaluate different approaches that combine a large generic translation model with domain-specific data. For this purpose, we use the sentences selected by the best approach ((5)$\wedge$(4)$\wedge$(3)) in the previous experiments and combine them with the generic parallel dataset. We evaluate the translation performance when *(i)* concatenating the selected domain-specific parallel dataset with the generic

Dataset Type	BLEU-2	BLEU-4	METEOR
Generic dataset	17.2	6.6	24.7
(5)∧(4)∧(3) sent. selec.	20.1	8.9*	27.2*
Data Concatenation (*i*)	18.1	6.8	24.1
Log-linear Models (*ii*)	18.9	8.1*	25.3
Fill-Up Model (*iii*)	17.7	7.0	24.7
(5)∧(4)∧(3) + IATE	19.8	9.0*	27.8*
(5)∧(4)∧(3) + DBpedia$^{(1)}$	20.6	9.1*	27.3*
(5)∧(4)∧(3) + DBpedia$^{(2)}$	**21.0**	**9.6**$^{*\diamond}$	**28.2**$^{*\diamond}$

Table 3: Evaluation of the ICD ontology evaluation dataset combining domain-specific with generic parallel knowledge and lexical enhancement of SMT using IATE and DBpedia (bold results = best performance; *statistically significant compared to baseline; $^{\diamond}$statistically significant compared to best sentence selection model)

parallel one, *(ii)* combining the generated translation models from the selected domain-specific parallel dataset and the generic corpus and *(iii)* applying the Fill-Up model to emphasise the domain-specific data in a single translation model. The translation performance of the combination methods are shown in Table 3. It is interesting to notice that none of them benefits from the use of the additional generic parallel data showing translation performance smaller than the domain-specific model. Although all methods outperform the generic translation model, only the log-linear approach, keeping in- and out-domain translation models separated, shows significant improvement. Comparing it to the combined sentence selection technique ((5)∧(4)∧(3)) does not show any statistical significant differences between the approaches. We conclude that the generic corpus is too large compared to the selected in-domain corpus, nullifying the influence of the extracted domain-specific parallel knowledge.

Lexical enhancement for SMT Since the out-of-vocabulary problem can be only mitigated with sentence selection, we accessed lexical resources IATE and DBpedia to further improve the translations of the medical labels. Based on the word overlap between labels and entries in IATE we extracted 11,641 English lexical entries with its equivalent in German. The DBpedia$^{(1)}$ approach, which disambiguates DBpedia entries based on the (Wikipedia article) categories (Arcan et al., 2012), identified 7,911 English-German expression for the targeted domain, while the abstract based disambiguation approach, marked as DBpedia$^{(2)}$ in Table 3 identified 3,791 bilingual entries. The lexical enhanced models further improved the translations of the medical labels (last three rows in Table 3) due to the additional bilingual information from the lexical resources, which is missing in the standalone sentence selection model. Comparing the ICD *evaluation dataset* and the translations generated with the DBpedia$^{(2)}$ lexical enhanced model we observed that more than 80 labels benefit from the additional lexical knowledge, e.g. correcting the mistranslated "adrenal gland" into "Nebenniere". The lexical extraction and disambiguation of bilingual knowledge based on the abstract of the article compared to the article categories further improves the lexical choice, helping SMT systems to improve the translation of ontology labels.

5.2 Manual Evaluation of Translated Labels

Since ontologies store specific vocabulary about a domain, this vocabulary is adapted to a concrete language and culture community (Cimiano et al., 2010). In order to investigate to what extent the automatically generated translations differ from a translator's adapted point of view, we manually inspected the translations produced by the sentence selection approaches described in Section 5.1.

While analysing the English and German part of the ICD ontology gold standard we noticed significant differences in the translations of the medical labels. As a result of the language and cultural adaptation, many labels in the ICD ontology were not always translated literally, i.e. parts of a label were semantically merged, omitted or new information was added while crossing the language border. For example, the ICD label "acute kidney failure and chronic kidney disease" is stored in the German part of the ontology as "Niereninsuffizienz".[16] Although none of the translation systems can generate the compounded medical expression for German, the SMT system generated nevertheless an acceptable translation, i.e. "akutes Nierenversagen und chronischer Nierenerkrankungen".[17] A more extreme example is the English label "slipping, tripping, stumbling and falls", in the German ICD ontology represented as

[16] Niereninsuffizienz←kidney insufficiency

[17] akutes←acute, Nierenversagen←kidney failure, und←and, chronischer←chronic, Nierenerkrankungen←kidney disease

”sonstige Stürze auf gleicher Ebene”.[18] The language and cultural adaptation is very active for this example, where the whole English label is semantically merged into the word ”Stürze”, meaning ”falls”. Additionally, the German part holds more information within the label, i.e. ”auf gleicher Ebene” (en. ”at the same level”), which is not represented on the English side. Since the SMT system will always try to translate every phrase (word or word segments) into the target language, an automatic translation evaluation cannot reflect the overall SMT performance.

Further we detected a large error class caused by compounding, a common linguistic feature of German. Although the phrase ”heart diseases” with its reference translation ”Herzkrankheiten” appears frequent in the generic training dataset, the SMT system prefers to translate it word by word into ”Herz Krankheiten”. [19] Similar observations were made with ”upper arm” (German ”Oberarm”) with the SMT word to word translation ”oberen Arm”.

Finally, we analysed the impact of the semantically enriched sentence selection with *related words* coming from Word2Vec compared to the surface based sentence selection, e.g. *preplexity*, *infrequent n-gram recovery* or *n-gram overlap*. Since semantically enriched selection stored the most relevant sentences, we observed the correct translation of the label ”blood vessels” into ”Blutgefäße”. The generic and other surface based selections translated the expression individually into ”Blut Schiffe”, where ”Schiffe” refers to the more common English word ”ship”, but not to ’part of the system transporting blood throughout our body’. The last example illustrates further the semantic mismatch between the training domain and the test domain. Using the generic model, built mainly out of European laws and parliament discussions (JRC-Acquis/Europarl) the word ”head” inside the label ”injury of head” is wrongly translated into the word ”Leiter”, meaning ”leader” in the legal domain. Nevertheless, the additional semantic information prevents storing wrong parallel sentences and guides the SMT to the correct translation, i.e. ”Schädigung des Kopfes”.[20]

[18] sonstige←other, Stürze←falls, auf←on, gleicher←same, Ebene←level

[19] Herz←heart, Krankheiten←diseases

[20] Schädigung←injury, des←of, Kopfes←head

6 Conclusion

In this paper we presented an approach to identify the most relevant sentences from a large generic parallel corpus, giving the possibility to translate highly specific ontology labels without particular in-domain parallel data. We enhanced furthermore the translation system build on the in-domain parallel knowledge with additional lexical knowledge accessing DBpedia. With the aim to better select relevant bilingual knowledge for SMT, we extend previous sentence and lexical selection techniques with additional semantic knowledge. Our proposed ontology-specific SMT system showed a statistical significant improvement (up to 3 BLEU points) of ontology label translation over the compared translation approaches.

In future, we plan to integrate a larger diversity of surface, semantic and linguistic information for relevant sentence selection. Although the SMT system is capable of translating several words into a compound word, the small amount of the selected sentences limits this capability. To improve the ontology label translations, we therefore see the need to focus more on the German compound feature. Additionally we observed that more than 25% of the identified lexical knowledge consists of multi-word-expressions, e.g. ”fatal familial insomnia”. For this reason, our ongoing work focuses on the alignment of nested knowledge inside those expressions. To move further in this direction, we plan to focus on exploiting morphological term variations taking advantage of the alternative terms provided by DBpedia.

Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight) and the European Union supported projects LIDER (ICT-2013.4.1-610782) and MixedEmotions (H2020-644632).

References

Arcan, M., Federmann, C., and Buitelaar, P. (2012). Experiments with term translation. In *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India.

Arcan, M., Giuliano, C., Turchi, M., and Buitelaar, P. (2014a). Identification of Bilingual Terms

from Monolingual Documents for Statistical Machine Translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, Dublin, Ireland.

Arcan, M., Turchi, M., Tonelli, S., and Buitelaar, P. (2014b). Enhancing statistical machine translation with bilingual terminology in a cat environment. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, Vancouver, Canada.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, Stroudsburg, PA, USA.

Bertoldi, N., Cettolo, M., and Federico, M. (2013). Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation. In *Proceedings of Machine Translation Summit XIV*, Nice, France.

Bicici, E. and Yuret, D. (2011). Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *Proceedings of IWSLT*.

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent Systems*, 14(1):20–26.

Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., and Gómez-Pérez, A. (2010). A note on ontology localization. *Appl. Ontol.*, 5(2):127–137.

Clark, J., Dyer, C., Lavie, A., and Smith, N. (2011). Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability . In *Proceedings of the Association for Computational Lingustics*.

Cuong, H. and Sima'an, K. (2014). Latent domain translation models in mix-of-domains haystack. In *Proceedings of the 25th International Conference on Computational Linguistics*, Dublin, Ireland.

Declerck, T., Pérez, A. G., Vela, O., Gantner, Z., Manzano, D., and D-Saarbrücken (2006). Multilingual lexical semantic resources for ontology translation. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Eck, M., Vogel, S., and Waibel, A. (2004). Language model adaptation for statistical machine translation based on information retrieval. In *Proc. of LREC*.

Erdmann, M., Nakayama, K., Hara, T., and Nishio, S. (2009). Improving the extraction of bilingual terminology from wikipedia. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5(4).

Espinoza, M., Montiel-Ponsoda, E., and Gómez-Pérez, A. (2009). Ontology localization. In *Proceedings of the Fifth International Conference on Knowledge Capture*, K-CAP '09, New York, NY, USA. ACM.

Fu, B., Brennan, R., and O'Sullivan, D. (2009). Cross-lingual ontology mapping - an investigation of the impact of machine translation. In Gómez-Pérez, A., Yu, Y., and Ding, Y., editors, *ASWC*, volume 5926 of *Lecture Notes in Computer Science*. Springer.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, Stroudsburg, PA, USA.

Gómez-Pérez, A., Vila-Suero, D., Montiel-Ponsoda, E., Gracia, J., and Aguado-de Cea, G. (2013). Guidelines for multilingual linked data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*. ACM.

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., and McCrae, J. (2012). Challenges for the multilingual web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11.

Haddow, B. and Koehn, P. (2012). Analysing the Effect of Out-of-Domain Data on SMT Systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montréal, Canada.

Harris, Z. (1954). Distributional structure. *Word*, 10(23).

Hildebrand, A. S., Eck, M., Vogel, S., and Waibel, A. (2005). Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT)*, Budapest.

Kirchhoff, K. and Bilmes, J. (2014). Submodularity for data selection in machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86. AAMT.

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., and Herbst, E. (2007). 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.

Langlais, P. (2002). Improving a general-purpose statistical translation engine by terminological lexicons. In *Proceedings of the 2nd International Workshop on Computational Terminology (COMPUTERM) '2002, Taipei, Taiwan*.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195.

Lü, Y., Huang, J., and Liu, Q. (2007). Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de Cea, G., and Cimiano, P. (2011). 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)*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, Stroudsburg, PA, USA.

Niehues, J. and Waibel, A. (2011). Using Wikipedia to Translate Domain-specific Terms in SMT. In *International Workshop on Spoken Language Translation*, San Francisco, CA, USA.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

O'Riain, S., Coughlan, B., Buitelaar, P., Declerck, T., Krieger, U., and Thomas, S. M. (2013). Cross-lingual querying and comparison of linked financial and business data. In Cimiano, P., Fernández, M., Lopez, V., Schlobach, S., and Völker, J., editors, *ESWC (Satellite Events)*, volume 7955 of *Lecture Notes in Computer Science*. Springer.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., and Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the Terminology and Knowledge Engineering (TKE2012) Conference*.

Ren, Z., Lü, Y., Cao, J., Liu, Q., and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, Stroudsburg, PA, USA.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). 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)*.

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Tiedemann, J. (2012). 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., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Tyers, F. M. and Pieanaar, J. A. (2008). Extracting bilingual word pairs from wikipedia. In *Collaboration: interoperability between people in the creation of language resources for less-resourced languages (A SALTMIL workshop)*.

Wu, H., Wang, H., and Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08.

Wuebker, J., Ney, H., Martínez-Villaronga, A., Giménez, A., , Juan, A., Servan, C., Dymetman, M., and Mirkin, S. (2014). Comparison of Data Selection Techniques for the Translation of Video Lectures. In *Proc. of the Eleventh Biennial Conf. of the Association for Machine Translation in the Americas (AMTA-2014)*, Vancouver (Canada).