Automatic Enrichment of Terminological Resources: the IATE RDF Example

Mihael Arcan[‡], Elena Montiel-Ponsoda[†], John P. McCrae[‡], Paul Buitelaar[‡]
[‡]Insight Centre for Data Analytics, National University of Ireland Galway
[firstname.lastname]@insight-centre.org
[†] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
emontiel@fi.upm.es

Abstract

Terminological resources have proven necessary in many organizations and institutions to ensure communication between experts. However, the maintenance of these resources is a very time-consuming and expensive process. Therefore, the work described in this contribution aims to automate the maintenance process of such resources. As an example, we demonstrate enriching the RDF version of IATE with new terms in the languages for which no translation was available, as well as with domain-disambiguated sentences and information about usage frequency. This is achieved by relying on machine translation trained on parallel corpora that contains the terms in question and multilingual word sense disambiguation performed on the context provided by the sentences. Our results show that for most languages translating the terms within a disambiguated context significantly outperforms the approach with randomly selected sentences.

Keywords: knowledge bases, terminology, multilinguality, machine translation

1. Motivation

Terminological resources have proven necessary in many organizations and institutions to ensure the quality and consistency of the terms used across the documents that serve expert communication. In the translation field, these resources are usually integrated into Computer Aided Translation (CAT) tools to allow translators to store the terms that appear in the documents they are translating. Terms and their translations may be accompanied by additional information intended to represent the meaning and the translation decisions taken for the use of the terms. The collected information accounts for the domain of the document, the definition of source or target terms, a sentence in which the term appears or other terms with the same meaning, i.e., term variants.

One of the most representative terminological resources is the Inter-Active Terminology for Europe (IATE).[1] It is a widely-used resource which incorporates all the terminology databases that had been independently built and maintained by the translation services of the different EU institutions. The public version of IATE contains approximately 1.4 million entries (8.6 million terms) in the 24 official EU languages. Due to the vast amount of domains covered in IATE (Table 1), the resource is a point of reference not only for EU translators and interpreters but also for translators and language service providers (LSPs) around the world. A simplified version of it is available in the TBX (Term Base eXchange) format for free download,[2] and has been converted to the Resource Description Framework (Klyne and Carroll, 2006, RDF) according to the linked data principles (Cimiano et al., 2015). As for any other language resource, the efforts that are required to maintain, update and clean the data are extensive. In addition to that, some languages or domains are extensively covered, whereas others have not received so much attention. Moreover, duplicate entries are common, since each institutional database was merged together into IATE in 2004. Furthermore, some terms are accompanied by definitions, contexts, and usage notes, and are supported by valid references and high reliability values, whereas others are not further described. Finally, due to the time pressure in the translation process, occasionally terminologists include terms or provide translations that are not properly checked, and thus not agreed or verified by experts.

Taking these issues into account, we claim that automatic processes should be put in place to support professionals in the enrichment, maintenance and cleaning of such a database, so that new terms in low covered languages could be added and enriched with contextual information, whereby duplicates, low quality or old terms could be removed. In order to contribute to the enrichment of IATE with terms that do not exist in some of the languages covered by the resource, we rely on Statistical Machine Translation (SMT) and leverage available translations in other languages, to identify relevant domain-specific sentences that contain those potential translations from a large set of parallel corpora. Furthermore, we use the identified relevant sentences to provide additional or missing contextual information to the terminological resource. Finally, we compare the usage of term variants in certain translations pairs, which can show a distinction between commonly used and less used terms in the IATE resource.

2. Related Work

Most of the previous work on the translation of knowledge resources, i.e., ontologies, taxonomies or terminological resources, 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 terms stored in the resource to be translated and the multilingual resource. Since the replacement of source terms with their translations within the dictionaries guarantees a high pre-

[1] http://iate.europa.eu/
[2] http://iate.europa.eu/tbxPageDownload.do

Language	Domain
Lang. Indep.	health (13,338), law (9,924), agriculture, forestry and fisheries (8,546), ...
Bulgarian	natural and applied sciences (2,397), eu institutions and european civil service (2,320), chemistry (2,294), ...
Czech	natural and applied sciences (608), health (601), finance (486), ...
Danish	health (66,777), agriculture, forestry and fisheries (42,195), natural and applied sciences (38,106), ...
German	law (50,139), land transport (29,801), executive power and public service (26,398), ...
Greek	no domain (7,735), communications (3,769), information technology and data processing (3,654), ...
English	health (28,543), information technology and data processing (28,063), ...
Spanish	law (957), no domain (927), politics (717), ...
Estonian	insurance (31), natural and applied sciences (29), agricultural activity (29), ...
Finnish	no domain (438), executive power and public service (306), agriculture, forestry and fisheries (280), ...
French	executive power and public service (11,685), natural environment (10,628), health (6,468), ...
Irish	natural and applied sciences (13), natural environment (7), agriculture, forestry and fisheries (2), ...
Croatian	fisheries (17), agriculture, forestry and fisheries (17), land transport (4), ...
Hungarian	executive power and public service (14), politics (4), political party (4), ...
Italian	no domain (1,262), natural environment (590), agriculture, forestry and fisheries (410), ...
Lithuanian	political party (7), culture and religion (6), agriculture, forestry and fisheries (5), ...
Latvian	politics (10), political party (9), regions of eu member states (5), ...
Maltese	health (4), finance (4), social protection (2), ...
Dutch	executive power and public service (1,955), humanities (992), education (951), ...
Polish	health (14), air and space transport (13), communications (11), ...
Portuguese	no domain (839), employment and working conditions (160), land transport (150), ...
Romanian	european union law (11), finance (8), land transport (5), ...
Slovak	health (12), chemistry (10), industry (8), ...
Slovene	no domain (12), political party (2), politics (2), ...
Slovak	health (6), law (6), executive power and public service (4), ...

Table 1: Statistics on most defined domains in IATE.

cision 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). BabelNet (Navigli, 2012), one of the largest multilingual knowledge bases, was created by linking Wikipedia entries and Wordnet synsets, and used commercial translation systems to fill in the missing lexical gap. Pérez and Berlanga (2015) enriched the non-English counterparts of the UMLS (Unified Medical Language System) knowledge resource in the biomedical domain. They used lexicons generated by word-alignment and machine translation approaches, and compared the results with the proposed semantics transfer approach, focusing on the semantic coherence of the generated translations between Spanish, French and German. Sajous et al. (2010) enriched Wiktionary by relying on similarity measures based on random walks through the graphs extracted from its lexical networks. In their final step they engaged users in collaborative editing in order to validate the resource. A different approach for translation and disambiguation of domain-specific expressions stored in knowledge bases was shown in Arcan et al. (2015), where the authors identified relevant in-domain parallel sentences and used them to train a small but domain-aware SMT system. Ordan et al. (2017) demonstrated an approach for bilingual dictionary creation using different translation directions within a loop. In contrast, de Melo and Weikum (2012) did not match concepts with SMT, but showed a machine learning approach which determines the best translation for English WordNet synsets by taking bilingual dictionaries, structural information of WordNet and corpus frequency information into account. Similarly, the multilingual disambiguation of ontology labels 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 and to link them across languages. Furthermore, McCrae et al. (2016) show positive effects of different domain adaptation techniques, i.e., using Web resources as additional bilingual knowledge, rescoring translations with Explicit Semantic Analysis (ESA) and language model adaptation for automatic knowledge base translation. To link knowledge graphs across languages, McCrae et al. (2017) propose a hybrid approach that combines dataset alignment and ontology translation techniques. The combination of these two techniques improves the translation of domain-specific expressions in comparison with approaches when used alone.

3. Methodology

We demonstrate an approach that uses the existing IATE terms to select the most relevant sentences in which those terms appear. By translating these sentences into languages for which no term is documented in the database, we obtain new translations for the available terms. Furthermore, we use these sentences to enrich IATE with additional information, i.e. how these terms appear in sentences of a representative domain as well as the usage of the translations in the parallel corpora.

Language	# of Entries	All Entries	Tokens	Types	Avg. Length	# of Unigrams
French	932,078	1,102,995	3,876,429	181,344	3.51	182,977
English	929,729	1,089,726	3,116,115	200,214	2.85	166,675
German	673,851	846,925	1,631,472	520,999	1.92	514,634
Italian	498,549	591,970	1,953,887	127,591	3.30	103,885
Dutch	489,941	590,319	1,412,233	311,212	2.39	288,864
Spanish	447,377	524,504	1,762,624	117,012	3.36	94,971
Danish	445,939	533,541	1,114,400	308,581	2.08	284,811
Greek	398,946	466,635	1,415,473	153,368	3.03	72,055
Portuguese	390,772	446,740	1,453,663	103,732	3.25	85,255
Finnish	251,544	298,639	546,823	198,219	1.83	163,764
Swedish	242,158	277,067	540,266	176,163	1.94	156,407
Irish	57,415	63,331	210,477	36,321	3.32	10,825
Polish	54,971	64,100	211,749	42,122	3.30	8,818
Slovene	43,168	49,181	150,681	32,127	3.06	8,781
Maltese	41,090	48,018	164,501	31,618	3.42	8,067
Lithuanian	38,281	43,799	134,902	28,426	3.08	6,046
Romanian	38,126	43,822	159,205	26,281	3.63	7,145
Estonian	34,957	43,735	100,196	37,303	2.29	17,406
Slovak	34,453	39,647	131,726	30,920	3.32	6,643
Hungarian	32,369	38,939	109,510	31,595	2.81	10,460
Bulgarian	31,960	36,856	135,502	24,456	3.67	5,313
Latvian	31,857	36,792	113,653	26,942	3.08	6,478
Czech	27,866	33,562	116,364	23,842	3.46	4,876
Croatian	14,635	16,440	55,139	15,003	3.35	2,317

Table 2: Statistics on covered terms in IATE.

Disambiguated Context Identification 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 (Gracia et al., 2012). Since manual multilingual translation and evaluation of knowledge bases is a very time-consuming and expensive process, we apply SMT to automatically translate domain-specific expressions and demonstrate its validity by translating the IATE entries. While an SMT system can only return the most frequent or dominant translation when given a term by itself, it has been showed that SMT provides strong word sense disambiguation when the word is given in the context of a sentence (Arcan et al., 2016a; Arcan et al., 2016b).

As a motivating example, we consider the word *vessel*, which appears several times in the IATE repository, whereby the most frequent translation into German is *Schiff*, with the meaning of 'a craft designed for water transportation', e.g., as given by Google Translate.[3] To overcome the issue of obtaining translations for *vessel* in other languages, and also in different domains (in the sense of *blood vessel*, for instance), we aim to identify (several) parallel sentences, which hold the terminological entries in the targeted domain, and use their context to translate them into other languages for which a translation does not exist. This means that if we know that the word *vessel* also represents the meaning of 'a tube in which a body fluid circulates' and the German translation for this entry is *Gefäß*, we look in our approach for sentences in a parallel corpus where the words *vessel* and *Gefäß* both occur and obtain a context such as 'blood vessel' that allows the SMT system to translate this entry correctly. Although a translation into German is not necessary in this case, since it is already documented in the database, we can use the English sentence (appropriately disambiguated) to translate the term into other languages, where the translation does not exist. To maximize our chances of finding a well-disambiguated sentence, we use existing terms in as many languages as possible.

Enhancing IATE with New Translations and Contextual Information In addition to the extension of IATE with missing translations for less covered languages, we further provide information on how the domain-specific expressions appear in sentences. Since we identified the relevant sentences by using linked IATE entries in different languages, we believe that this additional information can further enrich the terminological resource.

Frequency and Reliability of Term Variants When accessing IATE from its online interface,[4] it is common to find several translations for the same term in the same domain, the so-called term variants, and it is the user's decision to choose one variant over the others. In the same way, for each term separately, additional information about its usage or the level of confidence assigned by the creators of the resource can only be obtained in a time-consuming fashion. To overcome these issues, we use the data we obtain from the parallel corpus that is believed to represent the real use of terms.

[3] `https://translate.google.com`, September 2017

[4] `http://iate.europa.eu/SearchByQuery.do`

IATE Duplicates Due to the original merging of separately maintained databases, duplicates are common in IATE. To support the maintenance of the resource in a more efficient way, we evaluate how many IATE entries have identical terms in English and in other target languages.

4. Experimental Setting

In this section, we give an overview on the dataset and the translation toolkit used in our experiment and provide insights into the SMT evaluation techniques.

4.1. IATE - Inter-Active Terminology for Europe

IATE is the terminology database of the EU with its objective of supporting the EU translators and creating a terminology resource to ensure standardisation throughout all institutions. It incorporates the various terminology databases into one database containing approximately one million multilingual entries in English (Table 2).[5] Recent domains that have been extensively covered include the financial crisis, environment, fisheries and migration.

4.2. Statistical Machine Translation

Our approach is based on phrase-based SMT (Koehn et al., 2003), where we wish to find the best translation of a string, given by a log-linear model combining a set of features. The translation that maximizes the score of the log-linear model is obtained by searching all possible translation candidates. The decoder, which is a search procedure, provides the most probable translation based on a statistical translation model learned from the training data.

For our task, we use the statistical translation toolkit Moses (Koehn et al., 2007), where word alignments, necessary for generating translation models, were built with the GIZA++ toolkit (Och and Ney, 2003). The KenLM toolkit (Heafield, 2011) was used to build a 5-gram language model.

4.3. Parallel Resources for SMT training and Multilingual Word Sense Disambiguation

To ensure a broad lexical and domain coverage of our SMT system, we merged the existing parallel corpora for each language pair from the OPUS web page[6] into one parallel data set, i.e., Europarl (Koehn, 2005), DGT translation memories generated by the *Directorate-General for Translation* (Steinberger et al., 2014), MultiUN corpus (Eisele and Chen, 2010), EMEA, KDE4, OpenOffice (Tiedemann, 2009), OpenSubtitles2012 (Tiedemann, 2012). Similarly, we concatenated parallel corpora to identify relevant sentences containing IATE entries, which are then translated into the targeted languages. Table 3 shows the amount of parallel sentences used for the different language pairs.

4.4. Translation Evaluation Metrics

The automatic translation evaluation is based on the correspondence between the SMT output and reference translation (gold standard).

BLEU (Papineni et al., 2002) is calculated for individual translated segments (n-grams) by comparing them with a data set of reference translations. Considering the shortness of the entries in IATE, we report scores based on the unigram overlap (BLEU-1). Those scores, between 0 and 100 (perfect translation), are then averaged over the whole *evaluation data set* to reach an estimate of the translation's overall quality.

METEOR (Denkowski and Lavie, 2014) is based on the harmonic mean of precision and recall, whereby recall is weighted higher than precision. Along with exact word (or phrase) matching it uses additional features, i.e., stemming, paraphrasing and synonymy matching.

chrF3 (Popović, 2015) is a character n-gram metric, which has shown very good correlations with human judgements, especially when translating from English into morphologically rich languages (Stanojević et al., 2015).

The approximate randomization approach (Clark et al., 2011) is used to test whether differences among system performances are statistically significant.

		Tokens		Types	
Language Pair	Sent.	English	Target	English	Target
English-Bulgarian	33M	325M	279M	938k	1M
English-Czech	24M	278M	237M	1M	1M
English-Danish	17M	236M	212M	687k	1M
English-German	10M	145M	135M	561k	1M
English-Greek	34M	364M	330M	1M	1M
English-Spanish	37M	391M	378M	1M	1M
English-Estonian	15M	185M	144M	640k	1M
English-Finish	24M	293M	199M	826k	2M
English-French	53M	740M	795M	1M	1M
English-Irish	1M	14M	15M	180k	271k
English-Croatian	16M	165M	133M	626k	1M
English-Hungarian	36M	369M	298M	1M	3M
English-Italian	22M	269M	265M	934k	1M
English-Lithuanian	6M	108M	90M	421k	833k
English-Latvian	5M	102M	86M	389k	653k
English-Maltese	2M	40M	40M	218k	380k
English-Dutch	35M	394M	363M	976k	1M
English-Polish	34M	361M	295M	1M	1M
English-Portuguese	32M	369M	354M	1M	1M
English-Romanian	40M	385M	353M	1M	1M
English-Slovak	11M	144M	126M	543k	1M
English-Slovene	13M	161M	133M	631k	1M
English-Swedish	16M	193M	163M	687k	1M

Table 3: Statistics on parallel data for translation model training and word-sense disambiguation.

5. Results

In this section, we present the evaluation of the translated IATE entries into several languages not initially included in this resource, and how existing IATE terms have been exploited for our purposes in the parallel corpora used in this work.[7] In addition to this, we illustrate the enhancing of IATE RDF resource with additional contextual information

[5] Based on IATE TBX file - IATE_export_16032017.tbx

[6] http://opus.nlpl.eu/

[7] We randomly selected 2,000 terms, although not all target terms are represented in each language for evaluation.

	# of Terms		Random Context			Disambiguated Context		
	Evaluated	New	BLEU-1	METEOR	ChrF	BLEU-1	METEOR	ChrF
Bulgarian	406	1,594	72.3	36.5	86.3	78.5*	39.3	86.9
Danish	1,502	498	64.7	39.3	80.9	72.8*	43.5	84.4
Greek	1,289	711	61.8	52.3	81.4	69.3*	55.1	84.9
Spanish	1,504	496	77.7	37.4	77.3	84.2*	39.1	78.5
Finnish	1,014	986	45.2	26.4	75.2	51.1*	29.7	77.2
French	1,566	434	74.4	37.9	79.2	79.5*	39.2	79.6
Croatian	155	1,845	58.3	30.5	73.4	59.6	30.9	73.4
Italian	1,524	476	72.6	35.5	75.1	81.7*	38.3	77.6
Latvian	451	1,549	57.2	29.8	74.6	63.2*	33.6	79.5
Dutch	1,505	495	69.8	42.1	82.0	77.8*	46.6	86.1
Polish	566	1,434	53.8	25.1	70.2	63.4*	29.0	74.8
Portuguese	1,504	496	79.1	37.4	77.2	84.3*	38.9	78.4
Romanian	431	1,569	67.6	31.4	71.7	75.5*	35.5	74.7
Slovak	459	1,541	59.5	30.4	76.2	67.0*	34.7	79.4
Slovene	501	1,499	58.2	29.7	75.4	67.8*	34.3	79.2
Swedish	1,071	929	66.8	40.6	83.3	75.0*	45.7	85.9

Table 4: Automatic translation evaluation of IATE entries with random and disambiguated context (* statistically significant compared to the random context translation approach).

English	certification of products pursuant to specific airworthiness specifications , the related modifications , repairs and their continuing **airworthiness** , shall be charged as defined in tables 1 to 6 .
Italian	certificazione dei prodotti in conformità a determinate specifiche di aeronavigabilità , le relative modifiche , le riparazioni e la loro **aeronavigabilità** continua , sono contabilizzate come definito nelle tabelle da 1 a 6 .
Dutch	certificering van producten overeenkomstig specifieke luchtwaardigheidsspecificaties , de bijbehorende wijzigingen en reparaties en de permanente **luchtwaardigheid** daarvan , worden in rekening gebracht als omschreven in de tabellen 1 tot en met 6 .
Danish	certificering af produkter i henhold til de specifikke luftdygtighedsspecifikationer , de relaterede ændringer , reparationer og deres fortsatte **luftdygtighed** faktureres i overensstemmelse med tabel 1 til 6 .
Slovene	certifikacija proizvodov v skladu s posebnimi plovnostnimi specifikacijami , povezane spremembe , popravila in njihova stalna **plovnost** se zaračunavajo , kot je določeno v tabelah 1 do 6 .
Portuguese	a certificação de produtos em conformidade com especificações de aeronavegabilidade próprias , bem como as modificações e reparações associadas e respetiva **aeronavegabilidade** permanente , devem ser cobrados conforme definido nas tabelas 1 a 6

Table 5: Identified sentence for IATE entry *airworthiness*, and their translations in different languages.

(examples of sentences that show the real use of the term) and information about duplicates, so that they can be easily identified for an eventual cleaning of the resource.

5.1. Translation Evaluation

Table 4 illustrates the automatic translation evaluation for 16 languages. IATE entries are translated within a random and identified disambiguated context. Except for the Croatian language, translating IATE terminological expressions within a disambiguated context, statistically significantly (p-value < 0.01) outperforms the approach with randomly selected sentences. Due to this results, we believe that the newly added terms show similar translation quality.

5.2. Providing Disambiguated Contextual Information

To further enhance the IATE terminological resource, we believe that the identified disambiguated sentences can be beneficial for the users selecting a term due to the contextual information of the targeted domain. Therefore, we append the identified relevant sentences to the IATE RDF resource. Table 5 illustrates an example of the relevant sentence associated with the IATE term *airworthiness* (IATE-29309), with its translations in several languages.

5.3. Frequency and Reliability of Term Variants

With the aim of providing users with information about term variants and to help them choosing the best variant for their purposes, as well as to provide data to support the reliability score originally assigned by IATE terminologist, we perform an experiment on evaluating the appearance of terms in the parallel corpora used in this work. Table 6 illustrates examples of IATE terms in English that have more than one translation equivalent in the target languages, in this case, German, French and Slovene. These can be considered as term variants, and frequency numbers can give

IATE ID	English Term	En. Term Freq.	Target Term	S&T Freq.
IATE-913759	south caucasus	277	südkaukasus	229
			transkaukasien	1
IATE-1108878	body mass index	282	body-mass-index	126
			körpermasse-index	4
IATE-1126703	outward investment	122	investissement extérieur	7
			investissement réalisé à l'étranger	0
IATE-46262	electrical engineering	137	électrotechnique	36
			génie électrique	16
IATE-1211765	shovel	1,398	pelle	933
			bêche	10
	spade	645	pelle	66
			bêche	34
IATE-770947	state of the art	235	stanje tehnike	16
			najsodobnejša tehnologija	0
IATE-814939	distortive effect	42	izkrivljajoč učinek	7
			učinek izkrivljanja	6
	distorting effect	28	izkrivljajoč učinek	3
			učinek izkrivljanja	3

Table 6: Examples of IATE term frequencies in the parallel corpora.

us a clue on the real use of the term and, consequently, on how reliable it is to use it in a certain context.

Although the usage of *südkaukasus* and *transkaukasien* differs significantly based on the number of times these terms are mentioned in the parallel corpus, the IATE-assessed reliability score for both terms is the same (three out of four stars). Similarly, based on the parallel corpora, *body-mass-index* is highly preferred in comparison to the German translation *körpermasse-index*, but both terms again have same assessed reliability score (three out of four stars).

For the French translation of *outward investment*, the most used translation within the parallel corpus is *investissement extérieur*, whereby the additionally suggested term *investissement réalisé à l'étranger* documented in IATE does not appear in the corpus. On the other hand, the French translations of the term *electrical engineering*, i.e., *électrotechnique* and *génie électrique*, are both frequently mentioned in the used corpora. Furthermore, the entry *shovel* (IATE-1211765) is mostly aligned with the French term *pelle*, whereby *spade* is frequently translated as *pelle* as well as *bêche*.

For Slovene, *state of the art* is mostly aligned with *stanje tehnike* in the parallel corpus, whereby the additional term of the same IATE entry, *najsodobnejša tehnologija* is not a common translation for the English term according to the parallel corpus. The English terms *distortive effect* and *distorting effect*, both belonging to the IATE-814939 entry, are similarly frequent in English as well as their translations into Slovene.

5.4. IATE Duplicates

Table 7 presents the amount of duplicate English entries and their translations. As seen, most of the duplicate entries appear between the English and French language pair. As an example, *electronystagmography* in English and *électronystagmographie* in French are represented in two IATE entries, i.e., IATE-1289555 and IATE-1517532, although both entries have the same IATE `subjectField` (2841, i.e., *Medical science*). Similarly, the very specific IATE terms *international natural rubber council* in English and *internationaler Naturkautschukrat* in German belong to two separate IATE entries, i.e., IATE-151353 and IATE-777553. Different to the previous example, these two entries belong to different, but similar domains, i.e., *Industry* and *International trade*. As in the case of frequency of use and reliability, such a statistical corpus analysis would allow us to identify those cases in which entries have been duplicated, most probably because of the merging of similar terminological resources.

French	57,216	Finnish	9,677	Lithuanian	372
German	28,013	Slovak	8,293	Slovak	358
Italian	27,941	Irish	1,163	Estonian	352
Spanish	23,533	Maltese	683	Hungarian	314
Dutch	23,009	Slovene	519	Czech	296
Danish	22,331	Polish	503	Latvian	259
Portuguese	20,185	Romanian	489	Croatian	148
Greek	16,891	Bulgarian	373		

Table 7: Statistics on duplicates between IATE terms in English and their translations.

5.5. Publication

In order to maximize the availability of this data, this data is available under an open license (CC-BY) and was contributed to the RDF version of IATE (Cimiano et al., 2015) so that it will be part of the linked open data cloud. In order to distinguish this automatically created data from the existing manually created data in IATE, we used the PROV-O

ontology (Lebo et al., 2013).[8]

6. Conclusion and Future Work

In this work, we showed automatic approaches for maintaining and increasing the lexical coverage of knowledge bases, e.g. IATE. By identifying disambiguated context, we demonstrate statistical significant translation improvement for several languages. With this approach, we identified relevant sentences, which can be beneficial for users when short terms cannot be disambiguated without any context surrounding them. To differentiate commonly used and less preferred terms, we evaluate the usage of the IATE terms and their translations in parallel corpora. At last, we identify duplicates in IATE, which can help to maintain and clean up the terminological resource. As ongoing work we are focusing on neural machine translation to cross the language barrier and how to incorporate the lexical information as well as the semantic structure of the resource into an embedded space for translating domain-specific expressions (Arcan and Buitelaar, 2017). Furthermore, our focus will lie on existing terminological variations within IATE as well as variations provided by machine translation.

Acknowledgements

This publication is supported by a research grant from Science Foundation Ireland, SFI/12/RC/2289 (Insight), a research visit grant from Universidad Politécnica de Madrid, by the Spanish Datos4.0 project (TIN2016-78011-C4-4-R) and by the EU's Lynx project (H2020 Research and Innovation Programme under GA núm 780602).

7. Bibliographical References

Arcan, M. and Buitelaar, P. (2017). Translating domain-specific expressions in knowledge bases with neural machine translation. *arXiv preprint arXiv:1709.02184*.

Arcan, M., Turchi, M., and Buitelaar, P. (2015). Knowledge portability with semantic expansion of ontology labels. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, Beijing, China.

Arcan, M., Dragoni, M., and Buitelaar, P. (2016a). Translating ontologies in real-world settings. In *Proceedings of the 15th International Semantic Web Conference (ISWC-2016)*, Kobe, Japan.

Arcan, M., McCrae, J. P., and Buitelaar, P. (2016b). Expanding wordnets to new languages with multilingual sense disambiguation. In *26th International Conference on Computational Linguistics (COLING'16)*, Osaka, Japan.

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.

Cimiano, P., McCrae, J. P., Rodríguez-Doncel, V., Gornostay, T., Gómez-Pérez, A., Siemoneit, B., and Lagzdins, A. (2015). Linked terminologies: Applying linked data principles to terminological resources. In *Proceedings of the fourth biennial conference on electronic lexicography (eLex 2015)*, pages 504–517, Sussex, United Kingdom.

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*, Portland, Oregon.

de Melo, G. and Weikum, G. (2012). Constructing and utilizing wordnets using statistical methods. *Language Resources and Evaluation*, 46(2):287–311.

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

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*, Gothenburg, Sweden.

Eisele, A. and Chen, Y. (2010). MultiUN: A multilingual corpus from United Nation documents. In Daniel Tapias, et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5.

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

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

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.

Heafield, K. (2011). KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Klyne, G. and Carroll, J. J. (2006). Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation, World Wide Web Consortium.

Koehn, P., Och, F. J., and Marcu, D. (2003). 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*, pages 48–54, Edmonton, Canada. Association for Computational Linguistics.

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*

[8] for more information see `https://nuig.insight-centre.org/unlp/research/resources/multilingual-iate-rdf/`

Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Prague, Czech Republic.

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

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2013). PROV-O: The PROV ontology. W3c recommendation, World Wide Web Consortium.

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)*, Portland, Oregon.

McCrae, J. P., Arcan, M., Asooja, K., Gracia, J., Buitelaar, P., and Cimiano, P. (2016). Domain adaptation for ontology localization. *Web Semantics*, 36:23–31.

McCrae, J. P., Arcan, M., and Buitelaar, P. (2017). Linking knowledge graphs across languages with semantic similarity and machine translation. In *Workshop on Multi-Language Processing in a Globalising World (MLP2017)*, Dublin, Ireland.

Navigli, R. (2012). Babelnet goes to the (multilingual) semantic web. In *Proceedings of the 3rd Workshop on the Multilingual Semantic Web, Boston, USA, November 11th, 2012*.

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

Ordan, N., Gracia, J., and Kernerman, I. (2017). Autogenerating bilingual dictionaries. In *Proceedings of fifth biennial conference on electronic lexicography (eLex 2017),*, Leiden, Netherlands.

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.

Pérez, M. and Berlanga, R. (2015). Semantic transference for enriching multilingual biomedical knowledge resources. *Journal of biomedical informatics*, 58:1–10.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Sajous, F., Navarro, E., Gaume, B., Prévot, L., and Chudy, Y., (2010). *Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary*, pages 332–344. Springer Berlin Heidelberg, Berlin, Heidelberg.

Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. (2015). Results of the WMT15 Metrics Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, pages 256–273, Lisbon, Portugal, September.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., and Gilbro, S. (2014). An overview of the european union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.

Tiedemann, J. (2009). News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Advances in Natural Language Processing*, volume V, chapter V, pages 237–248. Borovets, Bulgaria.

Tiedemann, J. (2012). Character-based pivot translations for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 141–151, Avignon, France, April.