Enhancing Machine Translation of Academic Course Catalogues with Terminological Resources

Randy Scansani[1]
University of Bologna
Forlì, Italy

Silvia Bernardini[1]
University of Bologna
Forlì, Italy

Adriano Ferraresi[1]
University of Bologna
Forlì, Italy

Federico Gaspari[2]
Università per Stranieri "Dante Alighieri"
Reggio Calabria, Italy

Marcello Soffritti[1]
University of Bologna
Forlì, Italy

[1]name.surname@unibo.it, [2]gaspari@unistrada.it

Abstract

This paper describes an approach to translating course unit descriptions from Italian and German into English, using a phrase-based machine translation (MT) system. The genre is very prominent among those requiring translation by universities in European countries in which English is a non-native language. For each language combination, an in-domain bilingual corpus including course unit and degree program descriptions is used to train an MT engine, whose output is then compared to a baseline engine trained on the Europarl corpus. In a subsequent experiment, a bilingual terminology database is added to the training sets in both engines and its impact on the output quality is evaluated based on BLEU and post-editing score. Results suggest that the use of domain-specific corpora boosts the engines quality for both language combinations, especially for German-English, whereas adding terminological resources does not seem to bring notable benefits.

1 Introduction

1.1 Background

Following the Bologna process, universities have been urged to increase their degree of internationalization, with the aim of creating a European Higher Education Area (EHEA) that encourages students' mobility. This process has brought with it the need of communicating effectively in English also for institutions based in countries where this is not an official language. Nevertheless, previous work has shown that institutional academic communication has not undergone a substantial increase of translated content, both from a qualitative and from a quantitative point of view. Callahan and Herring (2012) claim that the number of universities whose website contents are translated into English varies across the European Union, with Northern and Western countries paying more attention to their internationalization than Southern ones. When quality is in focus, things do not improve: many of the translated documents feature terminological inconsistencies (Candel-Mora and Carrió-Pastor, 2014).

As one of the aims in the creation of the EHEA was to foster students' mobility, availability of multilingual course unit descriptions (or course catalogues) has become especially important. These texts start by indicating the faculty the course belongs to. After this, brief descriptions of the learning outcomes and of the course contents are given. The following sections outline the assessment and teaching methods. Lastly, details are provided regarding the number of ECTS credits for the course unit, useful links and readings for students, information about the lecturer's office hours and the language in which the course is taught.

Several aspects make these texts interesting for our purposes. First, they feature terms that are typical of institutional academic communication, but also expressions that belong to the discipline taught (Ferraresi, 2017). Second, they are usually drafted or translated by teachers and not by professional writers/translators (Fernandez Costales, 2012). Therefore, their disciplinary terminology is likely to be accurate, but they might not comply with the standards of institutional academic communication. Finally, they tend to be repetitive and relatively well-structured, and to be produced in large numbers on a yearly basis, through a mix of drafting from scratch and partial revisions or up-

The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT), pages 1–10,
Varna, Bulgaria, Sept 7, 2017
https://doi.org/10.26615/978-954-452-042-7_001

dates.

These characteristics make course catalogues an ideal test bed for the development of tools supporting translation and terminology harmonization in the institutional academic domain. Indeed, the development of such tools has been on the agenda of universities across Europe for several years now, as testified, e.g., by previous work in this area funded by the EU Commission in 2011 and involving ca. 10 European universities and private companies[1]. Despite its interest, this project does not seem to have undergone substantial development after 2013, nor does it seem to have had the desired impact on the community of stakeholders. In addition to that, it does not include one of our language combinations, i.e. Italian-English.

1.2 Objectives of the Study

In practical terms, being able to automatically translate texts that typically contain expressions belonging to different domains – the academic one and the disciplinary one – raises the question of how to choose the right resources and how to add them to the system in order to improve the output quality and to simplify post-editing. We aim at contributing to machine translation (MT) development not only understanding if MT results for translation in this domain are promising, but also finding out the most effective setup for MT engines, i.e. with a generic corpus, with an in-domain corpus or with one of these corpora and a bilingual glossary belonging to the educational or disciplinary domain.

In addition to focusing on developments for MT and its architecture, we are laying emphasis on MT contribution to the work of post-editors and to translation in the institutional academic domain. The development of an MT tool able to support professional and non-professional post-editors would speed up the translation of texts, thus favoring the internationalization of universities. Moreover, the present study is part of a larger project that aims to test the impact of terminology on output quality, post-editing effort and post-editor satisfaction[2]. Since terminology inconsistencies can negatively affect both output quality and post-editor's trust in an MT system, we are also investigating the relationship (if any) between the use of terminology resources at various stages of the MT-PE pipeline, and the perception of output quality and post-editing effort by professional and non-professional post-editors (Gaspari et al., 2014; Moorkens et al., 2015). To sum up, even if at these initial stages we are primarily interested in discovering the most effective architecture for our MT tool for this peculiar domain, we see these initial steps as crucially related to the overall application in a real-world scenario where human-machine interaction is of the essence.

For this study, a phrase-based statistical machine translation system (PBSMT) was used to translate course unit descriptions from Italian into English and from German into English. We built a baseline engine trained on a subset of the Europarl[3] corpus. Then, a small in-domain corpus including course unit descriptions and degree programs (see sect. 3.2) belonging to the disciplinary domain of the exact sciences was used to build our in-domain engine. We chose to limit our scope and concentrate on exact sciences since German and Italian degree programs whose course units belong to this domain translate their contents into English more often than other programs (the scarcity of high-quality human-translated parallel texts is arguably the major challenge for our work). We enriched the two training data sets with a bilingual terminology database belonging to the educational domain (see sect. 3.3) and we built two new engines: one trained on the Europarl corpus subset plus the bilingual terminology database, and one on the in-domain bilingual sentence pairs plus the bilingual terminology database. Each of the four engines for each language combination was then tuned on a subset of the in-domain corpus (more details about the resources are given in sect. 3). To evaluate the output quality, we are relying on two popular metrics: the widely-used BLEU score (Papineni et al., 2002) and post-editing score. The latter is based on edit-distance, like other popular methods such as TER or HTER (Snover et al., 2006), i.e. on the post-edits required to turn an MT output segment into its human reference. Even if our reference text is not a post-edited translation of the evaluation data set source side, we chose to also consider the PES results since they tend to be more clear for translators and post-editors.

[1] http://www.bologna-translation.eu/

[2] This work is part of a three-year project that will also include experiments with post-editors, aimed at measuring their reactions to machine-translated output enhanced with terminological backup information, as well as tests on neural machine translation (NMT).

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

2 Previous Work

A number of approaches have already been developed to use in-domain resources like corpora, terminology and multi-word expressions (MWEs) in statistical machine translation (SMT), to tackle the domain-adaptation challenge for MT. For example, the WMT 2007 shared task was focused on domain adaptation in a scenario in which a small in-domain corpus is available and has to be integrated with large generic corpora (Koehn and Schroeder, 2007; Civera and Juan, 2007). More recently, the work by Štajner et al. (2016) addressed the same problem and showed that an English-Portuguese PBSMT system in the IT domain achieved best results when trained on a large generic corpus and in-domain terminology.

Langlais (2002) showed that adding terminology to the phrase-table actually improved the WER score for the French-English combination in the military domain. For the same language combination, Bouamor et al. (2012) used pairs of MWEs extracted from the Europarl corpus as one of the training resources, but only observed a gain of 0.3% BLEU points (Papineni et al., 2002). Ren et al. (2009) extracted domain-specific MWEs from the training corpora showing encouraging improvements in terms of BLEU score for translations from English to Chinese in the patent domain. A sophisticated approach is the one described in Pinnis and Skadins (2012), where terms and named entities are extracted from in-domain corpora and then used as seeds to crawl the web and collect a comparable corpus from which more terms are extracted and then added to the training data. This method shows an improvement of up to 24.1% BLEU points for the English-Latvian combination in the automotive domain.

Methods to integrate terminology in MT have been recently developed focusing on how to dynamically insert terminology into a PBSMT system, i.e. injecting terminology in an MT engine without having to stop it or re-train it. Such methods suit the purpose of the present paper, as they focus (also) on Italian, German and English. Arcan et al. (2014a) tested for the first time the cache-based method (Bertoldi et al., 2013) to inject bilingual terms into a SMT system without having to stop it. This brought to an improvement of up to 15% BLEU score points for English-Italian in medical and IT domains. For the same domains and with the same languages (in both directions), Arcan et al. (2014b) developed an architecture to identify terminology in a source text and translate it using Wikipedia. This study resulted in an improvement of up to 13% BLEU score points. Moving to approaches focusing exclusively on morphologically complex languages, Pinnis (2015) reported on a new pre-processing method for the source text in order to extract terminology, translate it and add it to the MT system. An evaluation for English-German, English-Latvian and English-Lithuanian in the automotive domain showed an improvement of up to 3.41 BLEU points. The manual evaluation pointed out an increase of up to 52.6% in the number of the terms translated correctly.

3 Experimental Setup

3.1 Machine Translation System

The system we used to build the engines for this experiment is the open-source ModernMT (MMT)[4]. MMT (Bertoldi et al., 2017) is a project funded by the European Union which aims to provide translators and enterprises with a new and innovative phrase-based tool. The main reasons behind the choice of this software is that it is able to build custom engines without long and computationally complex training and tuning phases, providing high-quality translations for specific domains. As a matter of fact, recent evaluation (Bertoldi et al., 2017) carried out on texts from 8 different domains and for two language combinations (English-French and English-German), showed that MMT's training and tuning are faster than Moses', while their quality is similar. Besides, MMT outperforms Google Translate when translating texts belonging to specific domains on which the engine was trained.

For this work, we exploited both the tuning and testing procedures already implemented in MMT, i.e. a standard Minimum Error Rate Training (MERT) (Och, 2003) and a testing phase in which the engine can be evaluated on a specific set of data chosen by the user. The metrics used are the BLEU score (Papineni et al., 2002) and the post-editing score (PES), which is the inverse of the TER score (Snover et al., 2006).

3.2 Corpora

As mentioned in sect. 1.2, we enriched a baseline and an in-domain MT system using in-domain cor-

[4] http://www.modernmt.eu/

pora and terminology. Due to limitations of the computational resources available, training and tuning an engine on the whole Europarl corpus would not have been possible. We therefore extracted a subset of 300,000 sentence pairs from the Europarl corpus both for Italian-English and for German-English to use them as the training set of our baseline engine. Then, bilingual corpora belonging to the academic domain were needed as in-domain training, development and evaluation data sets for the two language combinations. Relying only on course unit descriptions to train our engines could have led to over-fitting of the models. Also, good-quality bilingual versions of course unit descriptions are often not available. To overcome these two issues we added a small number of degree program descriptions to our in-domain corpora, i.e. texts that are similar to course unit descriptions, but provide general information on a degree program, and are thus less focused on a specific discipline or course.

To build our in-domain corpora, we followed the method developed within the CODE project[5]. An Italian-English corpus of course catalogues was also available thanks to this project. The starting point for the search for institutional academic texts in German and Italian was the University Ranking provided by Webometrics[6]. This website ranks Higher Education Institutions from all over the world based on their web presence and impact. We crawled the top university websites in Italy and Germany and for each of the two countries we identified the four universities with the largest quantity of contents translated into English. From these, we downloaded texts within the exact sciences domain.

For the German-English combination, we collected two bigger corpora of course unit descriptions, and two smaller ones of degree program descriptions. For the Italian-English one we collected a corpus of course unit descriptions and two smaller corpora of degree program descriptions, to which we then added the CODE course unit descriptions corpus after cleaning of texts not belonging to the exact science domain. For both language combinations, each corpus was extracted from a different university website. We ended up with 35,200 segment pairs for German-English and 42,000 for Italian-English. The smallest corpus made of course unit descriptions was used as evaluation data set, i.e. 4,400 sentence pairs out of 35,200 for German-English and 3,700 out of 42,000 for Italian-English. 3,500 sentence pairs from the biggest course unit description corpus were extracted for each of the language combinations to exploit them as development set. The remaining sentence pairs – i.e. 27,300 for German-English and 34,800 for Italian-English – were used as training set for the in-domain engines.

3.3 Terminology

The terminology database was created merging three different IATE (InterActive Terminology for Europe)[7] termbases for both language pairs and adding to them terms and MWEs extracted from the Eurydice[8] glossaries. More specifically, the three different IATE termbases were the following: Education, Teaching, Organization of teaching. We also extracted the monolingual terms from the summary tables at the end of the five Eurydice volumes and we chose to keep just the terms in the fifth volume which are already translated into English and those terms whose translation in the target language was relatively straightforward (for example *Direttore di dipartimento* in Italian and "Head of department" in English).

The three different IATE files were in xml format and their terms were grouped based on their underlying concepts, so a single entry often contained more than one source term related to many target terms. For example one entry included the German terms *Erziehung und Unterricht* and *Unterricht und Erziehung*, that are translated into English as "educational services", "instruction services", "teaching" and "tuition". We extracted each term pair and merged them into a single plain-text (tab separated) bilingual termbase, where each pair has its own entry. We then collected the Eurydice bilingual terms and merged them with those extracted from IATE. At the end of this process, we obtained an Italian-English termbase with 4,143 bilingual entries and a German-English one with 5,465 bilingual entries.

In order to test the relevance of our term collec-

[5]CODE is a project aimed at building corpora and tools to support translation of course unit descriptions into English and drafting of these texts in English as a lingua franca. http://code.sslmit.unibo.it/doku.php

[6]http://www.webometrics.info/en

[7]http://iate.europa.eu/

[8]http://eacea.ec.europa.eu/education/eurydice/

tions for the experiment, we computed the number of types and tokens of the evaluation data set on the source side, and the number of termbase entries. We then compared these figures to the degree of overlap between the two resources, i.e. tokens and types occurring both in the termbase and in the source side of the evaluation data set for both language pairs, so as to gauge the relevance of the termbase. Since our termbase does not contain any inflected form, we are computing the degree of overlap only on the canonical form of the terms. Results are displayed in Tables 1 and 2.

It-En	
Corpus tokens	50,248
Corpus types	6,985
Termbase entries	4,142
Tokens overlap	20.44%
Types overlap	13.27%

Table 1: Types and tokens in the evaluation data sets, termbase entries, and type/token overlap between the two resources for It-En.

De-En	
Corpus tokens	26,956
Corpus types	5,614
Termbase entries	5,462
Tokens overlap	19.98%
Types overlap	9.63%

Table 2: Types and tokens in the evaluation data sets, termbase entries, and type/token overlap between the two resources for De-En.

The German-English corpus tokens are half those in the Italian-English corpus, while the Italian-English corpus includes ca. 1,300 types more than the German-English one (approximately 20% of the total number of Italian types). When German is the source language, the number of termbase entries is ca. one fifth of the corpus tokens, while the number of the Italian-English glossary entries is only one twelfth of the number of corpus tokens. The ratio between number of overlapping tokens and number of corpus tokens remains the same across the two language combinations (ca. 20%), while the ratio related to types is 13.27% for Italian-English and 9.63% for German-English. Based on these figures, we would expect our Italian-English termbase to have a slightly stronger influence on the output than the German-English one.

It is also interesting to observe the list of the glossary words occurring in the output ranked by their frequency for both source languages. For German the first five are *Informatik*, *Software*, *Vorlesung*, *Fakultät*, *Studium*, while for Italian we have: *corso*, *insegnamento*, *calcolo*, *prova*, *voto*. Considering the low degree of overlap and the large presence of basic words for the domain in both languages – and since most of them are unlikely to have multiple translations – we decided not to work on a time-demanding task such as solving the ambiguities in the termbase.

4 Experimental Results

For both language combinations we used MMT to create four engines:

- One engine trained on the subset of Europarl (baseline).
- One engine trained on the subset of Europarl and the terminology database (baseline+terms).
- One engine trained on the in-domain corpora (in-domain).
- One engine trained on the in-domain corpora and the terminology database (in-domain+terms).

Each engine was then tuned on a development set formed of 3,500 in-domain sentence pairs and evaluated on ca. 3,700 segment pairs (for Italian-English) and 4,400 segment couples (for German-English) (see sect. 3.2 for a description of the training, tuning and testing data sets).

4.1 Italian-English

For the engine that translates from Italian into English, results are shown in Table 3. If we compare the best in-domain engine to the best baseline according to the BLEU score, we can see that, after the tuning phase, the in-domain engine outperforms the baseline+terms by 7.85 points. Moreover, each engine has improved its performance according to both metrics after being tuned on our development set.

According to the automatic metrics, and contrary to our expectations, adding the terminology database did not influence the in-domain engines or the baseline ones in a substantial way,

sometimes actually causing a slight decrease in the engine performance. For example, our two in-domain engines had similar performance both before tuning – when their scores differed by 0.22 BLEU points and 0.19 PES points, with the in-domain outperforming its counterpart with terminology – and after tuning, when in-domain outperformed in-domain+terms by 0.78 BLEU points and 0.35 PES points.

To gain a slightly better insight into the engine performance, we quickly analyzed the outputs of the four engines. Many sentences contained untranslated words or word-order issues. The reference sentence "During the semester, two guided visits to relevant experimental workshops on the topics covered in the course will be organized" contained the words "semester" and "course" that are basic words of the domain and appear in the glossary. However, in both baseline engines the output is "During the half, will be organized two visits led to experimental laboratories relevant to the subjects raised during" and in both the in-domain ones the output is "During the semester, will be two visits to experimental laboratories pertinent to topics covered in the course". Two things are interesting here. First of all, the output confirms what the automatic metrics already highlighted: the output of the engines without terms is generally very similar to the output with terms. Moreover, adding the termbase did not substantially improve the generic output even when some of its words appeared in the termbase. We will discuss these results further in sect. 4.3.

It-En Engine	**BLEU**	**PES**
Baseline	16.90	35.27
Baseline tuned	22.58	40.08
Baseline+terms	17.09	35.04
Baseline+terms tuned	**22.75**	**40.36**
In-domain	26.72	50.61
In-domain tuned	**30.60**	**53.17**
In-domain+terms	26.50	50.42
In-domain+terms tuned	29.82	52.82

Table 3: Results for the Italian-English combination. For each engine, BLEU and PES scores are given both before and after tuning. The best baseline and in-domain results are shown in bold.

De-En Engine	**BLEU**	**PES**
Baseline	24.03	41.24
Baseline tuned	34.98	47.70
Baseline+terms	25.65	42.10
Baseline+terms tuned	**36.89**	**49.03**
In-domain	43.21	49.06
In-domain tuned	46.31	50.75
In-domain+terms	43.48	49.23
In-domain+terms tuned	**47.05**	**51.20**

Table 4: Results for the German-English combination. For each engine, BLEU and PES scores are given both before and after tuning. The best baseline and in-domain results are shown in bold.

4.2 German-English

Table 4 shows results for the German-English language combination. After tuning, the best in-domain engine outperformed the baseline by 10.16 points according to BLEU and by 2.17 points according to PES. The tuning performed on the in-domain engines causes an improvement of more than 3% in terms of BLEU score. Regarding the baseline engines, the tuning enhances the quality by ca. 10 BLEU points and 7 PES points, thus narrowing the performance gap between the two different kinds of engines.

What is important to notice is that, counter-intuitively and similarly to what we observed for the Italian-English combination, the collection of academic terminology does not affect the translation output quality: the metrics show an improvement of 0.74 BLEU points and 0.45 PES points when terminology is added to the in-domain engine (results after the tuning phase). The addition of terminology seems to be slightly more effective on the baseline engines, improving the automatic scores by 1.91 BLEU points and 1.33 PES points after tuning.

A quick analysis of the output shows the same issues identified in sect. 4.1. One example is the reference sentence "Lecture, exercises, programming exercises for individual study", that is translated as "Lecture, exercises, Programmieraufgaben zum private study" in both in-domain engines and as "Lecture, exercise, Programmieraufgaben on Selbststudium" in both baseline engines (the word couple *Vorlesung*-Lecture was in the termbase). The same German word was not translated in some sentences of the two in-domain engines outputs – e.g. "Vorlesung (presentation of

Slides and presentation interactive examples)", for the engine with terms and "Vorlesung (presentation of presentation and interaktiver examples)" for the engine without terms – while it was in the baseline ones: "Lecture (presentation of Folien and idea interactive examples)". This is another negative result for our termbase, since neither of the in-domain engines translated "Vorlesung" as "lecture", while the baseline ones did without the help of the terminological resource. We will further discuss these results in sect. 4.3.

4.3 Discussion

In our experimental setup, adding terminology to a general-purpose engine and to an in-domain engine does not influence the output quality substantially. We compared the figures in Tables 1 and 2 (regarding the degree of overlap between the evaluation data set and the bilingual glossary) to the automatic scores assigned to our engines (Tables 3 and 4) to investigate the impact on output quality. The degree of tokens overlap between the bilingual glossary and the evaluation data set is similar for the two source languages (ca. 20%). Despite this, for the German-English combination the baseline+terms engine outperformed the baseline engine by 1.91 BLEU points and 1.33 PES points, which is the largest gain obtained in this study adding a bilingual glossary to the training data set. If we look at the baseline and baseline+terms engines for Italian-English, for example, the latter outperformed the former by only 0.17 BLEU and 0.28 PES points. This might suggest that the target terms in the German-English glossary were consistent with those used in the reference text, while for Italian-English there were more discrepancies between the two resources.

Another variable that has to be taken into account is the way in which terms are extracted and injected into the MT engine. As reported in sect. 2, methods in which terminology is extracted from other resources and then added to training data sets (Bouamor et al., 2012) are less effective than, for example, approaches in which terminology is extracted from the training data set (Ren et al., 2009; Pinnis and Skadinš, 2012) or injected dynamically, i.e. at run-time without re-training, into the MT engine (Arcan et al., 2014a). In our case the Italian-English term pairs were 4,143 against the 34,800 sentence pairs of the training data set, while for German-English we had 5,465 term pairs as compared to 27,300 sentence pairs. Due to the difference in the amount of term pairs and segment pairs, simply adding the glossary to the sentence pairs might cause it to lose its influence on the training process.

If we look at Tables 3 and 4, we can see that in-domain segments boost the engine quality both during training – with the in-domain engines outperforming the baseline – and after tuning, which brings remarkable improvements. This could suggest that the PBSMT system is able to extract academic terms and expressions from the in-domain corpus, without the need of being enhanced with a termbase belonging to the same domain. Additional evidence for this are the examples in section 4.1, where some of the termbase words were not translated in the baseline engines output of both language combinations, while they were correctly translated in both in-domain engines. As a matter of fact, the terminology database was able to increase the score by more than 1% in terms of BLEU only on one occasion – i.e. the baseline engine for German-English –, while for the other three engine pairs (in-domain with and without terms for German-English, baseline and in-domain with and without terms for Italian-English) the performance increased by few decimals or even decreased. The same happens if we look at the PES score.

To further discuss the results without using the BLEU metric, whose figures are often less intuitive than the PES' ones especially for translators and post-editors, it is interesting to notice how the in-domain engines for both language combinations always reach at least 50% in terms of PES score after tuning. Despite the low quality of the examples seen in sections 4.1 and 4.2, these PES scores are an encouraging result if we consider that we are carrying out the first experiment on this domain and that we are exploiting quite a small amount of in-domain resources to build our engine, a condition that is likely to remain constant given the nature of communication in this domain. It also suggests that, in this domain, MT is likely to boost the post-editor's productivity if compared to a translation from scratch. Moreover, we expect to obtain further improvements building an engine combining both generic and in-domain resources in the training phase, so as to hopefully observe a further increase of the PES and hence of the post-editor productivity.

5 Conclusion and Further Work

This paper has described an attempt at evaluating the potential of the use of in-domain resources (terminology and corpora) with MT in the institutional academic domain, and more precisely for the translation of course unit descriptions from German into English and from Italian into English. Following the results of the present experiment, we are planning to carry out further work in this field.

Since academic terms are only one subset of the terminology used in course unit descriptions, which also includes terms related to the subject matter of each unit, it would be interesting to investigate the advantages of adding disciplinary terminology alongside the academic one. We therefore plan to combine academic and disciplinary terminology. Following the encouraging results of the baseline engine tuned on the in-domain resources, we also plan to investigate the performance of an engine trained on both generic and in-domain resources and tuned on an in-domain development set. As shown in the work by Štajner et al. (2016), PBSMT systems' performance increases when the training data set includes a small quantity of in-domain resources – corpora or termbase – and a large generic corpus.

As we have seen in section 3.3 the overlap afforded by the termbase used for this experiment was less than optimal and its structure would require an accurate procedure to extract the most likely term pair for this domain, since a source term often has multiple target translations. For these reasons and basing on the results, IATE is probably not the best resource for our purposes. As part of future work, we are interested in extracting terminology from other resources, e.g. the UCL-K.U.Leuven University Terminology Database[9], or, ideally, from a resource developed collaboratively by universities across Europe, consistent with EU-wide terminological efforts but more readily usable and focusing on agreed-upon terms and with limited ambiguity. We will also test methods to make available the most relevant terms for the texts to be translated, i.e. extracting terminology from the training data. In both cases – use of external resources or extraction from the training data set – we are planning to add inflected forms of the terms. In addition to their extraction, we are considering injection of terms into an MT system in other ways than simply adding them to the training set, where the termbase is likely to play a minor role because of its small size compared to the corpora. In future work we will compare methods to add terms at run-time in a post-editing environment, in order to analyze the impact of these suggestions on the post-editors' work. What we are expecting from this experiment is to find a way to increase the post-editor trust in the suggested terminology, and hence in the MT engine.

As this was our first attempt to build an MT engine in this domain, sometimes we were forced to concentrate on more technical aspects, e.g. the improvements in the BLEU score to analyze the engines' development. In future work we are planning to use metrics that better take into account terminology translation (e.g. precision, recall, f-measure) and also manual evaluation to collect more data on the impact of our work on the post-editing phase.

To conclude, in this paper we have taken a first step toward the development of a tool that combines machine translation, corpora and terminology databases, with the aim of streamlining the provision of course unit descriptions in English by European universities. Our in-domain engines showed encouraging results, even if – as expected – they are not able to boost a post-editor's productivity yet, while the role of terminology (what kind, how it is injected into the engine) is still to be further investigated, as is the confidence-building potential of quality terminology databases on post-editing work.

[9] https://sites.uclouvain.be/lexique/lexique.php

Acknowledgments

The authors would like to thank Mauro Cettolo, Marcello Federico and Luisa Bentivogli of Fondazione Bruno Kessler (FBK) for their advice and for help with ModernMT, and three anonymous reviews for insightful comments on the first draft of this paper. The usual disclaimers apply.

References

Mihael Arcan, Claudio Giuliano, Marco Turchi, and Paul Buitelaar. 2014b. Identification of bilingual terms from monolingual documents for statistical machine translation. In *Proceedings of the 4th International Workshop on Computa-*

tional Terminology. Dublin, Ireland, pages 22–31. http://www.aclweb.org/anthology/W14-4803.

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2014a. Enhancing statistical machine translation with bilingual terminology in a CAT environment. In Yaser Al-Onaizan and Michel Simard, editors, *Proceedings of AMTA 2014*. Vancouver, BC.

Nicola Bertoldi, Roldano Cattoni, Mauro Cettolo, Amin Farajian, Marcello Federico, Davide Caroselli, Luca Mastrostefano, Andrea Rossi, Marco Trombetti, Ulrich Germann, and David Madl. 2017. MMT: New open source MT for the translation industry. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*. Prague, pages 86–91. https://ufal.mff.cuni.cz/eamt2017/user-project-product-papers/papers/user/EAMT2017_paper_88.pdf.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In Andy Way, Khalil Sima'an, Mikel L. Forcada, Daniel Grasmick, and Heidi Depaetere, editors, *Proceedings of the XIV Machine Translation Summit*. Nice, France, pages 35–42.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, pages 674–679. ACL Anthology Identifier: L12-1527. http://www.lrec-conf.org/proceedings/lrec2012/pdf/886_Paper.pdf.

Ewa Callahan and Susan C. Herring. 2012. Language choice on university websites: Longitudinal trends. *Journal of International Communication* 6 (2012):322–355.

Miguel Ángel Candel-Mora and María Luisa Carrió-Pastor. 2014. Terminology standardization strategies towards the consolidation of the European Higher Education Area. *Procedia - Social and Behavioral Sciences* 116:166 – 171.

Jorge Civera and Alfons Juan. 2007. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic, pages 177–180. http://www.aclweb.org/anthology/W/W07/W07-0222.

Alberto Fernandez Costales. 2012. The internationalization of institutional websites. In Anthony Pym and David Orrego-Carmona, editors, *Translation Research Projects*. Tarragona: Intercultural Studies Group, pages 51–60.

Adriano Ferraresi. 2017. Terminology in European university settings. The case of course unit descriptions. In Paola Faini, editor, *Terminological Approaches in the European Context*. Cambridge Scholars Publishing, Newcastle upon Tyne, pages 20–40.

Federico Gaspari, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. Perception vs reality: Measuring machine translation post-editing productivity. In Sharon O'Brien, Michel Simard, and Lucia Specia, editors, *Proceedings of AMTA 2014*. Vancouver, BC, pages 60–72.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, StatMT '07, pages 224–227. http://dl.acm.org/citation.cfm?id=1626355.1626388.

Philippe Langlais. 2002. Improving a general-purpose statistical translation engine by terminological lexicons. In *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology - Volume 14*. Association for Computational Linguistics, Stroudsburg, PA, USA, COMPUTERM '02, pages 1–7. https://doi.org/10.3115/1118771.1118776.

Joss Moorkens, Sharon O'Brien, Igor A. L. da Silva, Norma B. de Lima Fonseca, and Fábio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3-4):267–284.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '03, pages 160–167. https://doi.org/10.3115/1075096.1075117.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Mārcis Pinnis. 2015. Dynamic terminology integration methods in statistical machine translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*. Antalya, Turkey, pages 89–96.

Mārcis Pinnis and Raivis Skadinš. 2012. MT adaptation for under-resourced domains - what works and what not. In *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012*. Tartu, Estonia, pages 176–184.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*. Association for Computational Linguistics, Suntec, Singapore, MWE '09, pages 47–54.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*. Cambridge, Massachussets, pages 223–231.

Sanja Štajner, Andreia Querido, Nuno Rendeiro, João António Rodrigues, and António Branco. 2016. Use of domain-specific language resources in machine translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France, pages 592–598.