Assessing the Use of Terminology in Phrase-Based Statistical Machine Translation for Academic Course Catalogues Translation

Randy Scansani
University of Bologna
Forlì, Italy
randy.scansani@unibo.it

Marcello Federico
Fondazione Bruno Kessler
Trento, Italy
federico@fbk.eu

Luisa Bentivogli
Fondazione Bruno Kessler
Trento, Italy
bentivo@fbk.eu

Abstract

English. In this contribution we describe an approach to evaluate the use of terminology in a phrase-based machine translation system to translate course unit descriptions from Italian into English. The genre is very prominent among those requiring translation by universities in European countries where English is not a native language. Two MT engines are trained on an in-domain bilingual corpus and a subset of the Europarl corpus, and one of them is enhanced adding a bilingual termbase to its training data. Overall systems' performance is assessed through the BLEU score, whereas the f-score is used to focus the evaluation on term translation. Furthermore, a manual analysis of the terms is carried out. Results suggest that in some cases - despite the simplistic approach implemented to inject terms into the MT system - the termbase was able to bias the word choice of the engine.

Italiano. *Nel presente lavoro viene descritto un metodo per valutare l'uso di terminologia in un sistema PBSMT per tradurre descrizioni di unità formative dall'italiano in inglese. La traduzione di questo genere di testi è fondamentale per le università di Paesi europei dove l'inglese non è una lingua ufficiale. Due sistemi di MT vengono addestrati su un corpus in-domain e un sottoinsieme del corpus Europarl. Ad uno dei due sistemi viene aggiunto un glossario bilingue. La valutazione delle prestazioni globali dei sistemi avviene tramite BLEU score, mentre f-score usato per la valutazione specifica della traduzione dei termini. È stata inoltre condotta un'analisi manuale dei termini. I risultati evidenziano che, nonostante il metodo elementare utilizzato per inserire i termini nel sistema di MT, il termbase in alcuni casi in grado di infuenzare la scelta dei termini nell'output.*

1 Introduction

Availability of *course unit descriptions* or *course catalogues* in multiple languages has started to play a key role for universities especially after the Bologna process (European Commission et al., 2015) and the resulting growth in student mobility. These texts aim at providing students with all the relevant information regarding contents, prerequisites, learning outcomes, etc.

Since course unit descriptions have to be drafted in large quantities on a yearly basis, universities would benefit from the use of machine translation (MT). Indeed, the importance of developing MT tools in this domain is further testified by two previous projects funded by the EU Commission, i.e. TraMOOC[1] and Bologna Translation Service[2]. The former differs from the present work since it does not focus on academic courses, while the latter does not seem to have undergone substantial development after 2013 and in addition to that, it does not include the Italian-English language combination.

Automatically producing multilingual versions of course unit descriptions poses a number of challenges. A first major issue for MT systems is the scarcity of high quality human-translated parallel texts of course unit descriptions. Also, descriptions feature not only terms that are typical of institutional academic communication, but also expressions that belong to specific disciplines (Ferraresi, 2017). This makes it cumbersome to

[1] Translation for Massive Open Online Course http://tramooc.eu/

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

choose the right resources and the most effective method to add them to the MT engine.

For this study, we chose to concentrate on course units belonging to the disciplinary domain of exact sciences, since Italian degree programmes whose course units belong to this domain translate their contents into English more often than other programmes.

A phrase-based statistical machine translation system (PBSMT) was used to translate course unit descriptions from Italian into English. We trained one engine on a subset of the Europarl corpus and on a small in-domain corpus including course unit descriptions and degree programs (see sect. 3.1) belonging to the domain of the exact sciences. Then, we enriched the training data set with a bilingual terminology database belonging to the educational domain (see sect. 3.2) and built a new engine. To assess the overall performance of the two systems we automatically evaluated them with the BLEU score. We then focused on the evaluation of terminology translation, by computing the f-score on the list of termbase entries occurring both in the system outputs and in the reference translation (see sect. 4). Finally, to gather more information on term translation, a manual analysis was carried out (see sect. 5).

2 Previous work

A number of approaches have already been developed to use in-domain resources like corpora and terminology in statistical machine translation (SMT), indirectly tackling 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). Recently, the work by Štajner et al. (2016) 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.

For French-English in the military domain, Langlais (2002) reported on improvements of the WER score after using existing terminological resources as constraints to reduce the search space. For the same language combination, Bouamor et al. (2012) used couples of MWEs extracted from the Europarl corpus as one of the training resources, yet only observing a gain of 0.3% BLEU points (Papineni et al., 2002).

Other experiments have focused on how to insert terms in an MT system without having to stop or re-train it. These dynamic methods suit the purpose of the present paper, as they focus (also) on Italian-English. Arcan et al. (2014b) injected bilingual terms into a SMT system dynamically, observing an improvement of up to 15% BLEU points for English-Italian in medical and IT domains. For the same domains and with the same languages (in both directions), Arcan et al. (2014a) developed an architecture to identify terminology in a source text and translate it using Wikipedia as a resource. The terms obtained were then dynamically added to the SMT system. This study resulted in an improvement of up to 13% BLEU score points.

We have seen that results for the languages we are working on are encouraging, but since they are strongly influenced by several factors – i.e. the domain and the injection method – an experiment on academic institutional texts is required in order to test the influence of bilingual terminology resources on the output.

3 Experimental Setup

3.1 Corpora

A subset of 300,000 sentence pairs was extracted from the Europarl Italian-English bilingual corpus (Koehn, 2005). Limiting the number of sentence pairs of the generic corpus was necessary due to limitations of the computational resources available. Then, bilingual corpora belonging to the academic domain were needed as development and evaluation data sets and to enhance the training data set. One course unit description corpus was available thanks to the CODE project[3]. After cleaning of texts not belonging to the exact science domain, we merged the corpus with other two smaller corpora made of course unit descriptions. We then extracted 3,500 sentence pairs to use them as development set.

Relying only on course unit descriptions to train our engines could have led to an over-fitting of the models. Moreover, high quality parallel course unit descriptions are often difficult to be found. To

[3] 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

Data Set	Sent. pairs	It Tokens	En Tokens
Training (Europarl)	300,000	7,848,936	8,046,827
Training (in-domain)	34,800	441,030	399,395
Development	3,500	48,671	43,919
Test	3,465	49,066	45,595

Table 1: Number of sentence pairs and tokens in each of the data sets used.

overcome these two issues we added a small number of degree program descriptions to our in-domain corpus. To conclude, a fourth small course unit descriptions corpus was built to be used as evaluation data set. All the details regarding the sentence pairs and tokens are provided in Table 1.

3.2 Terminology

The terminology database was created merging three different IATE (InterActive Terminology for Europe)[4] termbases for both languages and adding to them the terms extracted from the fifth volume of the Eurydice[5] glossaries. More specifically, the three different IATE termbases were: Education, Teaching, Organization of teaching.

To verify the relevance of our termbase with respect to the training data we measured its coverage. Since the terms in the termbase are in their base form, in order to obtain a more accurate estimate we lemmatised[6] the training sets before calculating the overlap between the two resources.

As we can see in Table 2, the 24.08% of the termbase entries are also in the source side of the two training corpora, and 29.19% are in the target side, meaning that the two resources complement each other well.

	It	En
Europarl lemmas	7,848,936	8,046,827
In-domain lemmas	441,030	399,395
Termbase entries	4,142	4,142
Europarl overlap	23.03%	29.20%
In-domain overlap	27.52%	29.33%
Total overlap	24.08%	29.19%

Table 2: Number of lemmas in the generic and in-domain training sets, termbase entries, and coverage of the termbase wrt. training data.

3.3 Machine Translation System

We tested the performance of a PBSMT system trained on the resources described in sections 3.1 and 3.2. The system used to build the engines for this experiment is the open-source ModernMT (MMT)[7] (Bertoldi et al., 2017). Two engines were built in MMT:

- One engine trained on the subset of Europarl plus our in-domain corpus.
- One engine trained on the subset of Europarl plus our in-domain corpus and the terminology database.

Both engines were tuned on our development set and evaluated on the test set (see sect. 3.1).

4 Experimental results

To provide information on the overall translation quality of our PBSMT engines, we calculated the BLEU scores (Papineni et al., 2002) obtained on the test set. Table 3 shows the results for both engines, where the engine without terminology is referred to as *w/o terms* and the one with terminology is referred to as *w/ terms*.

Furthermore, we evaluated the systems focusing on their performance on terminology translation. To this purpose, we relied on the f-score. More in detail, for both engines we extracted the number of English termbase entries appearing in the system output and in the reference translation. Exploiting these figures, we were able to compute Precision, Recall and f-score. Results are reported in Table 4.

Engine	BLEU
w/o terms	25.92
w/ terms	26.00

Table 3: BLEU score for the two engines.

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

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

[6] Lemmatisation was performed using the TreeTagger: https://goo.gl/JjHMcZ

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

	w/o terms	**w/ terms**
Terms in ref	1,133	1,133
Terms in output	1,061	1,083
Correct terms	633	630
Precision	0.596	0.581
Recall	0.558	0.555
F-score	0.577	0.568

Table 4: Number of occurrences of termbase entries in the reference and in the output texts, number of terms in the reference appearing also in the outputs, Precision, Recall and F-score.

The figures in Tables 3 and 4 show that adding our termbase to the training data set does not affect the output in a substantial way. While according to the BLEU score the w/ terms engine slightly outperforms the w/o terms engine, the f-score – indicating performance on term translation – is marginally higher for the w/o terms system.

Focusing on the usage of terminology, a number of observations can be made. As regards the distribution of termbase entries in the test set - which contains 3,465 sentence pairs - it is interesting to know that the number of output and reference sentences containing at least one term is fairly low, i.e. 945 (27.30%) for the reference text, 866 (24.99%) for the w/o terms output and 870 (25.10%) for the w/ terms output.

Considering the terms found in the two outputs, we observe that their number only differs by 23 units (ca. 2% of the number of terms in the outputs). Also, the number of overlapping terms is very high, i.e. 882 terms (out of 1,061 for the engine w/o terms and out of 1,083 for the engine w/ terms). As a matter of fact, the top-6 frequent terms in the systems' outputs are the same – *course*, *oral*, *ability*, *lecture*, *technology* and *teacher* – and cover approximately a half of the total amount of extracted terms for both outputs.

We then compared the English termbase entries appearing in the target side of the test set to those appearing in the training set. Each of the 78 terms occurring at least one time in the test set (corresponding to 1,133 total occurrences as reported in Table 4), also occur in the training set – out of which 60 in its in-domain component.

However, even though our training data cover the total amount of terms present in the test data, and despite the high overlap between the terms produced by the two engines, still there is a considerable number of terms that are different. We thus cannot exclude an influence of the termbase on the word choice of the w/ terms system. For this reason, an in-depth analysis of the different terms produced by the two engines was carried out.

5 Manual Evaluation

The analysis of the sentences where the termbase entries used by the two engines differed showed that in some cases the termbase forced the system to use its target term even if a different translation - sometimes also correct - was present in the training corpora. Some examples are reported in Table 5. For the source words *prova orale* (Example 1) and *esame scritto* (Example 2), the engine w/ terms used *oral examination* and *written examination*, while the one w/o terms used *written exam* and *oral exam*, but only the occurrences with *examination* are in the termbase. Moreover, Example 2 also includes the termbase word *preparazione*, which is translated with *preparation* by the engine w/ terms, while it is not translated at all by the engine w/o terms.

Another interesting example is the translation of the source word *docente* (Example 3), where the termbase corrected a wrong translation. The Italian term was wrongly translated with *lecture* by the engine w/o terminology, and with *teacher* - which is the right translation for this text - by the engine w/ terminology.

In Example 4, the Italian sentence contained the termbase entry *voto finale*, which was translated with *final vote* by the engine w/o terms and with the termbase MWE *final mark* by the w/ terms engine. Also in this case the termbase corrected a mistake, since *vote* is not the correct translation of *voto* in this context.

The comparison between the two engines' outputs shows that, even though our training data covered the total amount of terms present in the test set, the termbase influenced the MT output of the engine w/ terms biasing the weights assigned to a specific translation.

Such results have to be judged taking into account the preliminary nature of this study, aimed at understanding the practical implications of using terminology in PBSMT, and therefore exploiting a simplistic approach to inject terms. As a matter of fact, we found that also some of the termbase entries occurring in the reference – e.g. *certifica-*

SRC	La **prova orale** si svolgerà sugli argomenti del programma del corso.	
REF	The **oral verification** will be on the topics of the lectures.	
W/O TERMS	The **oral exam** will take place on the program of the course.	✓
W/ TERMS	The **oral examination** will take place on the program of the course.	✓
SRC	La **preparazione** dello studente sarà valutata in un **esame scritto**.	
REF	Student **preparation** shall be evaluated by a 3 hrs **written examination**.	
W/O TERMS	The student will be evaluated in a **written exam**.	✗
W/ TERMS	The **preparation** of the student will be evaluated in a **written examination**.	✓
SRC	Ogni **docente** titolare	
REF	Each **lecturer**.	
W/O TERMS	Every **lecture**.	✗
W/ TERMS	Every **teacher**.	✓
SRC	In tal caso il **voto finale** terrà conto anche della prova orale.	
REF	In this case the **final score** will be based also on the oral part.	
W/O TERMS	In this case the **final vote** will take account the oral test.	✗
W/ TERMS	In this case the **final mark** will be based also the oral test.	✓

Table 5: MT output examples showing the influence of the termbase on the word choice of the w/terms engine. Note that the ✓ and ✗ marks refer to human assessment and not to the correspondence with the reference.

tion, *instructor*, *text book*, *educational material* – were not used in the output of the system w/ terms and this is probably due to the limitations of our method. The terms *instructor*, *text book* and *educational material* did not occur in the w/o terms output neither, while *certification* did.

To sum up, what emerges is that using terminology in PBSMT to translate course catalogues - and more specifically course unit descriptions - can influence the MT output. In our case, since the improvements were measured against the output of the w/o terms engine - which might eventually be correct even if using different terms from those included in the termbase - the metrics results were not informative enough and a manual analysis of the terms had to be carried out.

6 Conclusion and further work

This paper has described a preliminary analysis aimed at assessing the use of in-domain terminology in PBSMT in the institutional academic domain, and more precisely for the translation of course unit descriptions from Italian into English. Following the results of the present experiment and given its preliminary nature, we are planning to carry out further work in this field.

In section 4 we have seen that the institutional academic terms contained in our testing data also appeared in the training data, thus limiting the impact of terminology on the output. However, course catalogues and course unit descriptions include terms belonging to the specific disciplines (see sect. 1) as well. In our future works we are therefore planning to focus not only on academic terminology, but also on the disciplinary one testing its impact on the output of an MT engine translating course unit descriptions.

After this first experiment on the widely-used PBSMT architecture, in future work we are planning to exploit neural machine translation (NMT). In particular, our goal is to develop an NMT engine able to handle terminology correctly in this text domain, in order to investigate its effect on the post-editor's work. For this reason, a termbase focused on the institutional academic domain, e.g. the UCL-K.U.Leuven University Terminology Database[8] or the Innsbrucker Termbank 2.0[9] could be used to select an adequate benchmark for the development and evaluation of an MT engine with a high degree of accuracy in the translation of terms.

Ackowledgements

The authors would like to thank Silvia Bernardini, Marcello Soffritti and Adriano Ferraresi from the University of Bologna for their advice on terminology and institutional academic communication, and Mauro Cettolo from FBK for help with ModernMT. The usual disclaimers apply.

[8] https://goo.gl/huoevR

[9] https://goo.gl/W2GH5h

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 Computational 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.

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.

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.

European Commission, EACEA, and Eurydice. 2015. *The European Higher Education Area in 2015: Bologna Process Implementation Report*. Luxembourg: Publications office of the European Union.

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.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT, Phuket, Thailand, pages 79–86. http://mt-archive.info/MTS-2005-Koehn.pdf.

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.

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.

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.