MAGMATic: A Multi-domain Academic Gold Standard with Manual Annotation of Terminology for Machine Translation Evaluation

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

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

Silvia Bernardini
University of Bologna
Forlì, Italy
silvia.bernardini@unibo.it

Adriano Ferraresi
University of Bologna
Forlì, Italy
adriano.ferraresi@unibo.it

Abstract

This paper presents MAGMATic (Multi-domain Academic Gold Standard with Manual Annotation of Terminology), a novel Italian–English benchmark which allows MT evaluation focused on terminology translation. The data set comprises 2,056 parallel sentences extracted from institutional academic texts, namely course unit and degree program descriptions. This text type is particularly interesting since it contains terminology from multiple domains, e.g. education and different academic disciplines described in the texts. All terms in the English target side of the data set were manually identified and annotated with a domain label, for a total of 7,517 annotated terms. Due to their peculiar features, institutional academic texts represent an interesting test bed for MT. As a further contribution of this paper, we investigate the feasibility of exploiting MT for the translation of this type of documents. To this aim, we evaluate two state-of-the-art Neural MT systems on MAGMATic, focusing on their ability to translate domain-specific terminology.

1 Introduction

The availability of bilingual versions of course catalogues has started to play a major role for European universities after the Bologna Process and the resulting growth in students' mobility. Course catalogues fall into the category of institutional academic text collections and they usually include degree program and course unit descriptions, where information regarding degree courses and modules are provided to students. Such texts have to be produced and published every year in each country language and in English. Universities would thus undoubtedly benefit from the use of machine translation (MT).

Further proof of the need for an MT engine able to translate course catalogues are two projects funded by the European Commission, namely the Bologna Translation Service[1] (Depraetere et al., 2011), aimed at developing an MT system to translate course catalogues in 9 language combinations, and TraMOOC,[2] aimed at using MT for the translation of massive online open courses from English into eleven European and BRIC languages.

Developing an engine in this field poses several challenges. First, the fact that degree program and course unit descriptions are usually translated by non-native speakers of the target language (Fernandez Costales, 2012) reduces the number of available high-quality and alignable bilingual texts. Moreover, the lack of guidelines and best practices to draft these texts results in substantial unmotivated variation among course catalogues from different universities. Finally, institutional academic texts usually contain terminology from different domains, with disciplinary terms, e.g. *Hydrosilylation*, *Fotoredox catalysis*, for a course on chemistry, appearing together with educational ones - e.g. *ECTS*, *module*.

The potential and challenges mentioned so far make course catalogues an interesting test bed for

[1] https://cordis.europa.eu/project/rcn/191739/factsheet/en
[2] http://tramooc.eu/content/scientific-publications

neural MT (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014). Indeed, in the last few years NMT has delivered considerable improvements in output quality in many respects (Bentivogli et al., 2016), yet not showing clear-cut progresses when it comes to lexis-related issues, e.g. lexical choices, omissions or mistranslations (Castilho et al., 2018). These issues are especially critical for texts rich in domain-specific terminology, or texts containing terms belonging to different domains. Testing an MT engine on course catalogues can provide interesting information on domain-specific terminology handling and on results achievable with a relatively small amount of in-domain resources used to perform domain-adaptation of a neural model.

Whilst assessing systems' ability to correctly translate domain-specific terms is a crucial aspect in MT evaluation, research in the field has to cope with a dearth of publicly available resources specifically tailored to that task. The main contribution of this paper is to provide the MT community with MAGMATic (Multi-domain Academic Gold standard with Manual Annotation of Terminology), a novel Italian–English benchmark which allows MT evaluation focused on terminology translation. The data set comprises 2,056 sentences extracted from course unit and degree program descriptions from four different Italian universities and manually aligned to their English translations. All terms in the English target side of the data set were manually identified and annotated with a domain label, for a total of 7,517 annotated terms, covering 20 different domains related to different disciplines - excluding humanities and with a focus on hard sciences - as well as education and education equipment. These features make the data set a valuable resource to evaluate and analyze systems' performance on terminology translation, thus contributing to shed light on this crucial aspect for MT. MAGMATic is released under a Creative Commons Attribution – Non Commercial – Share Alike 4.0 International license (CC BY-NC–SA 4.0), and is freely downloadable at:

`https://ict.fbk.eu/magmatic/`

In the remainder of this paper we describe MAGMATic and illustrate its potential by using it to evaluate two state-of-the-art MT systems (Google Translate and ModernMT), both in terms of overall performance and focusing on their ability to translate domain-specific terminology. After describing related work on term translation evaluation (Section 2), we introduce the main characteristics of MAGMATic (Section 3) and provide results from the evaluation study carried out on the two state-of-the-art MT systems (Section 4).

2 Related work

A number of monolingual annotated data sets for benchmarking terminology extraction and classification techniques have been created along the years for different domains (Kim et al., 2003; Bernier-Colborne and Drouin, 2014; Q. Zadeh and Handschuh, 2014; Astrakhantsev et al., 2015). The situation is much less favourable for terminology translation evaluation. Indeed, the majority of works addressing domain adaptation for MT evaluate systems only in terms of overall performance on a domain-specific test set, while very few studies specifically focus on the engines' ability to translate domain-specific terminology, and thus resort to test sets in which terms are annotated. To the best of our knowledge, only the following manually annotated resources are made available to the community. The BitterCorpus[3] (Arcan et al., 2014a) is a collection of parallel English–Italian documents in the information technology domain in which technical terms in both the source and target sides of the bi-texts are manually marked and aligned. TermTraGS[4] (Farajian et al., 2018) is a sentence-aligned version of the BitterCorpus, which also includes a large training set.

On a different aspect of MT quality evaluation, most of the works comparing NMT with previous paradigms treat correct or wrong lexical choices as one of the main quality indicators (Bentivogli et al., 2016; Bentivogli et al., 2018; Toral and Sánchez-Cartagena, 2017; Castilho et al., 2018; Van Brussel et al., 2018). However, all these works focus on the broader concept of lexical issues without specifically addressing terminology. The MAGMATic data set offers a new opportunity to compare different MT approaches directly on terminology issues.

Finally, regarding the institutional academic scenario, it is worthwhile to point out that neither of the two EU-funded projects mentioned in Sect. 1 – *Bologna Translation Service* and *TraMOOC* – led to the creation of data sets targeted to the evaluation of terminology translation.

[3] `https://ict.fbk.eu/bittercorpus/`
[4] `https://gitlab.com/farajian/TermTraGS`

3 Data set description

3.1 Data selection

The text material used in this work was collected from the websites of four Italian universities. All the course unit and degree program descriptions for which the corresponding English version was available were extracted, automatically aligned at sentence level and cleaned with TMop (Jalili Sabet et al., 2016), an open-source software for Translation Memory cleaning.

As an attempt to narrow down the number of domains – and thus the variability of terminology – course catalogues belonging to the humanities and social sciences were excluded, keeping only those catalogues related to scientific disciplines.

Then, a subset of sentence pairs was randomly selected and manually checked to ensure alignment correctness. This procedure resulted in 2,157 Italian–English parallel sentences. Statistics for the data set are summarised in Table 1.

MAGMATic		
	It	**En**
Sent.pairs	2,157	
Tokens	36,162	34,589
Vocabulary	10,207	9,138

Table 1: Size of the MAGMATic data set: number of sentences, number of tokens (i.e. running words) and vocabulary (i.e. number of distinct word types).

3.2 Data annotation

Two expert linguists with a background in translation studies took part in the annotation: one of them annotated the whole data set and the other annotated a portion of it so as to allow inter-annotator agreement assessment (see details in Section 3.4).

Two main annotation tasks were performed on the English target side of the data set, namely *(i)* the identification of the terms and *(ii)* their classification into domain categories. In order to ensure annotation quality and comparability, guidelines were created, tested in a pilot study and then given to the annotators.

Term identification. Both single-word (SW) terms – i.e. terms formed of one word – and multi-word (MW) terms – i.e. terms formed of two or more words – were annotated.

Furthermore, instances of language for general and specific purposes often blur into each other, making the decision as to what belongs to one or the other prone to subjectivity bias. For this reason, annotators were asked to report on their level of confidence, distinguishing between `sure` terms and `possible` terms, the latter accounting for expressions whose terminological status and specialisation were uncertain. For example, in a description of a course on electronics, *RC-circuit* was identified as a `sure` term and *charge* as a `possible` term. Where contents of a course on chemistry were outlined, *analysis* was categorized as `possible` and *pollutants formation* as `sure`. In sentences describing a course's teaching and evaluation methods, *exam* and *lecture* were labelled as `sure` terms, while *topics* and *notions* were labelled as `possible`. This additional annotation level is particularly useful since it supports more flexible evaluation designs.

Domain annotation. The identified terms were assigned to one of the following categories:

- `Disciplinary`: the term belongs to a disciplinary domain - e.g. *chemical reaction, linear equation, cholinesterase*.
- `Education`: the term belongs to the educational domain - e.g. *module, course, lecturer*.
- `Education equipment`: the term refers to educational equipment that could also be used elsewhere - e.g. *overhead projector, desk*.

While the `education` and `education equipment` categories are univocal, the `disciplinary` category encompasses multiple domains, i.e. multiple scientific disciplines. To assign each term to a specific discipline, we leveraged the names of the degree programs included in the data set: each sentence in the data set was automatically labelled with its corresponding degree program name and all the terms annotated as `disciplinary` in those sentences during the annotation process inherited the sentence domain label by default. Annotators were shown this domain label during the annotation process and asked to signal cases where a discrepancy between the label assigned automatically and the actual domain of one or more terms was observed. In these cases, annotators were asked to manually assign a different label to the term(s), selecting it from the list of degree program names.

The annotation was carried out using the MT-EQuAL annotation tool (Girardi et al., 2014). For

	Disciplinary		Education		Equipment		Total
	Sure	Poss.	Sure	Poss.	Sure	Poss.	
SWs	2,298	295	868	323	111	21	**3,916**
MWs	2,464	359	491	186	85	16	**3,601**
Total	4,762	654	1,359	509	196	37	
	5,416		**1,868**		**233**		**7,517**
Vocabulary	**4,316**		**686**		**130**		**5,132**

Table 2: Statistics of the terms annotated in the MAGMATic data sets. Terms in the three domain categories - Disciplinary, Education, Education-equipment (here Equip.) - are further split into the Sure and Possible (Poss.) subcategories. For either of these subcategories, the number of SWs and MWs, and the total number of terms are provided. In the two bottom rows, the total number of terms and the vocabulary (i.e. the number of distinct terms) are given for each category.

each English sentence, the MT-EQuAL interface displays the source sentence and the disciplinary domain retrieved from the name of the university course catalogue. Furthermore, the tool allows the annotators to perform the two annotation steps simultaneously: they mark each term and annotate it (`sure`/`possible` distinction and domain category) in a single go. This makes the annotation task efficient and less demanding in terms of effort.

3.3 Annotation statistics

Details regarding the number of terms annotated in the data set are provided in Table 2. In 101 sentences out of 2,157 (see Table 1) no terms were found. We therefore ended up with 2,056 sentence pairs and a total of 7,517 term occurrences, which correspond to 5,132 distinct terms.

The `disciplinary` category is the largest, while the `education equipment` category is the smallest. Looking at the proportion between `sure` and `possible` terms for each category, it is interesting to note that `possible` terms are much more frequent in the `education` category (27.2% of the total terms) than in the `disciplinary` (12%) or `education equipment` (15.9%) categories. We can assume that `disciplinary` or `education equipment` terms are rarely encountered in everyday language, and are thus easier to identify as terms. On the other hand, `education`-related terms are also used outside of the domain, making the decision as to their status more difficult.

Looking at SWs and MWs, their number in the data set is approximately the same. However the `disciplinary` category contains more MWs than SWs, whereas for the two other categories the opposite is the case. This is in line with what was stated above, i.e. `disciplinary` terms are highly domain-specific, and thus more likely to be MWs than, for example, `education` ones. The average length of MW terms is 2.44 words.

Comparing the number of term occurrences with the corresponding vocabulary, we see that terms in the `education` category show a much lower degree of variation than `disciplinary` terms. Indeed, the type-token ratio amounts to 0.80 for the `disciplinary` category, 0.37 for `education` and 0.56 for `education equipment`. This is due to the fact that the `disciplinary` category includes multiple domains, and thus a high number of different terms, while `education` and `education equipment` terms are stable and repeated across most texts. Also, the 5 most frequent terms in the data set belong to the `education` category (SWs: *student, course, students, knowledge, lectures*; MWs: *oral exam, end of the course, written test, oral examination, written exam*).

As concerns the specific domains represented in the `disciplinary` category, we saw in the previous section that the specific domain labels were assigned to the terms by exploiting the names of the degree programs of the universities from which the data set was derived. These names refer to domains with different granularity - e.g. biology, which is more generic, and biotechnology, which is more specific - and thus different size. To obtain a more homogeneous set of domains, we merged the most specific ones with the generic ones where appropriate, e.g. biotechnology was grouped with biology and biomedicine with medicine. This procedure resulted in 20 macro-domain labels with a similar level of granularity.

Examples of the macro-domains are given in Table 3, which shows the 5 most and 5 less populated ones. As we can see in the table, the number of terms included in the most populated domains al-

low for an extremely thorough terminology evaluation. Also, even if not all of them are displayed here, 9 domains out of 20 include more than 300 annotated terms. Regarding the less populated domains, they appear frequently in translation tasks and only three of them contain less than 100 annotated terms.

Domain	SWs	MWs	Total
Chemistry	345	367	712
Informatics	256	224	480
Physics	184	283	467
Biology	245	212	457
Mechanical engineering	200	210	410
...	...	...	...
Geosciences	62	47	109
Industrial engineering	48	59	107
Astronomy	21	61	82
Law	15	34	49
Institutions	14	11	25

Table 3: The 5 most populated and 5 least populated macro-domains covered in the data set and number of terms in each of them (SW, MW and total).

3.4 Inter-annotator agreement

In order to assess the reliability of the annotations, 220 sentences – corresponding to 10% of the data set – were annotated by a second annotator.

Inter-Annotator Agreement (IAA) was calculated for the two types of manual annotation, namely *(i)* the identification of the terms and *(ii)* their assignment to a domain category.

Agreement was computed on all the identified terms, without taking into account the `sure`/`possible` distinction.

Term identification. Two different types of agreement were calculated, to account for *complete* as well as *partial* agreement. Complete agreement refers to perfect overlap of two terms annotated by different annotators (i.e. exact match), whereas for partial agreement overlap is calculated at the level of the single words composing the term.

The agreement rates – computed using the Dice coefficient[5] (Dice, 1945) – are 0.69 for complete agreement and 0.79 for partial agreement. Given the high number of MW terms and the strict approach used for complete agreement, results may be considered satisfactory in terms of reliability of the annotations and suitability of the annotation guidelines.

Domain annotation. For the subset of terms for which complete agreement between the two annotators was found (495 terms), we also calculated the agreement on the assigned category label (i.e. `disciplinary`, `education`, `education equipment`).

To this end, we computed the standard *kappa coefficient* κ (in Scott's π formulation) (Scott, 1955; Artstein and Poesio, 2008), which measures the agreement between two raters, each of whom classifies N items into C mutually exclusive categories, taking into account the agreement occurring by chance.

The resulting κ value is 0.95, which – according to the standard interpretation of the κ values (Landis and Koch, 1977) – corresponds to "almost perfect" agreement.

4 MT evaluation on MAGMATic

4.1 MT and institutional academic texts

As a first application of our MAGMATic data set, we evaluated translations of course catalogues produced by two state-of-the-art NMT systems, i.e. Google Translate (GT)[6] and ModernMT (MMT)[7].

On the one hand, course catalogues are an ideal test bed for MT, given the multi-domain nature of these texts. On the other hand, being able to apply MT to course catalogues is particularly key for universities, since the increasing students and staff mobility has created the need of translating a large quantity of institutional academic texts into English (see Sect. 1).

Given the lack of in-house (customised) MT systems and of high-quality in-domain parallel data, using such technologies is a big challenge for higher-education institutions. Two ready-to-use state-of-the-art MT systems like MMT and GT thus represent a viable solution for this real-world multi-domain translation scenario. Both of them are based on the state-of-the-art transformer architecture (Vaswani et al., 2017) and trained on a large pool of parallel data. Furthermore, MMT implements an adaption mechanism which allows the system to adapt to new data in real time (Bertoldi

[5]Note that Dice coefficient has the same value of the F1 measure computed considering either annotator as the reference.

[6]`translate.google.com`

[7]`www.modernmt.eu`

et al., 2018). This feature represents a particularly interesting option in our scenario, since it would allow universities to leverage new translated data as soon as they are produced. In our evaluation we used the full-fledged commercial version of MMT available through the MateCat tool[8] and and we compared it with the GT online system.[9]

To the best of our knowledge, this contribution represents the first attempt at translating institutional academic texts with NMT.

4.2 Evaluation scenarios

Given the novelty of the application of MT to the translation of course catalogues, we are focusing on two scenarios that we deem realistic for one or more universities willing to use MT:

- First scenario (GT, MMT-I). One or more universities want to use MT for the translation of their course catalogues for the first time, and have no translation memories. At this point, no in-domain bilingual texts are available.
- Second scenario (GT, MMT-II). A university consortium agrees to coordinate their communication strategies. They use CAT tools for translating their course catalogues and produce a reasonable amount of translations, which can be leveraged as shared domain-adaptation data.

In order to address the second scenario, we needed an in-domain data set to be exploited for MT adaptation. To this effect, the parallel data collected from the 4 Italian universities but left out in the creation of MAGMATic (see Sect. 3.1) were used. Statistics for this data set are outlined in Table 4.[10]

Since the online generic version of GT used in this work is not adaptive, it can be tested in the first evaluation scenario only. As a SOTA system, GT provides an external validation of the quality of MMT. Differently, MMT is evaluated in both scenarios to analyse the impact of in-domain data on translation quality.

4.3 Evaluation metrics

The MT systems were evaluated both in terms of overall performance and specifically targeting their ability to translate domain terminology.

[8] www.matecat.com

[9] Evaluations were carried out on February 5th, 2019.

[10] The statistics for MAGMATic, which was used as test set, are shown in Table 1.

Domain-adaptation		
	It	**En**
Sent.pairs	40,361	
Tokens	632,223	601,236
Vocabulary	55,458	48,126

Table 4: Size of the domain-adaptation data set: number of sentences, number of tokens (i.e. running words) and vocabulary (i.e. number of distinct word types).

The bigger picture of the quality achieved with the setup described so far is provided through an automatic evaluation in terms of BLEU score (Papineni et al., 2002).

The evaluation focused on terminology translation is based on the Term Hit Rate (THR) metric (Farajian et al., 2018). THR takes in a list of annotated terms in each reference sentence and looks for their occurrence in the MT output. Then it computes the proportion of terms in the reference that are correctly translated by the MT system. An upper bound of 1 match for each reference term is applied in order not to reward over-generated terms in the MT output.

Similarly to the approach adopted for inter-annotator agreement (see Sect. 3.4), two THR types are computed: *perfect THR* – where a match is scored only if the whole reference term appears in the MT output – and *partial THR*, where the overlap between the reference terms and the MT output is calculated at the level of shared tokens. In this case, function words are removed from the MW terms in the reference, so as to avoid false positives with other function words present in the MT output.

	BLEU (↑)
GT	36.90
MMT-I	35.45
MMT-II	43.16

Table 5: BLEU scores for GT and for MMT in both scenarios.

4.4 Evaluation results

A general overview on the quality achieved by GT, MMT-I (first scenario) and MMT-II (second scenario) is provided in Table 5.

The good results obtained by GT and MMT-I show that NMT can be helpful already in the first scenario, where only generic systems can be used. The huge performance increase of MMT-II (+7.71

Perfect THR									
	GT			MMT-I			MMT-II		
	Overall	SWs	MWs	Overall	SWs	MWs	Overall	SWs	MWs
All	63.72	75.43	50.98	60.97	72.98	47.90	65.33	76.07	53.65
Disc	66.80	79.75	54.91	63.94	77.52	51.47	67.74	80.03	56.50
Edu	55.62	66.33	36.78	53.32	63.48	35.45	59.28	68.01	44.61
Equip	55.78	66.96	36.76	53.31	64.10	34.96	59.11	68.40	43.32
Sure	64.95	76.26	52.76	62.43	73.91	50.06	66.58	77.05	55.30
Poss	57.25	71.20	41.35	53.25	68.23	36.18	58.75	71.05	44.74

Table 6: Perfect THR for GT and the 2 MMT systems. In addition to the overall scores, figures for SWs and MWs are given separately. Results are provided *(i)* for the whole data set (All), *(ii)* split according to the domain category (Disc, Edu, Equip) and *(iii)* distinguishing between `sure` and `possible` terms.

wrt MMT-I and +6.26 wrt GT) is even more encouraging in the long-term perspective.

Focusing on the evaluation of terminology translation, perfect and partial THR scores were computed on MAGMATic for GT and the two MMT systems.

Table 6 presents results for Perfect THR. Since MAGMATic contains both SW and MW terms, the table gives the scores for each set separately in addition to the overall score. Also, to allow a more detailed analysis of the systems' behaviour on MAGMATic terms, results are provided by domain category (`disciplinary`, `education`, `equipment`) and in terms of the `sure`/`possible` distinction.

Considering the strict parameters used to calculate perfect THR, the results shown in Table 6 are quite satisfactory. Regarding domain categories, all systems in all scenarios perform far better on `disciplinary` terms. As for term length, SW terms are, as expected, easier to translate than MWs. The most challenging terms for all MT systems are MWs in the `education` and `equipment` categories.

Focusing on the first scenario, we see that GT and MMT have a similar behaviour, since the differences between the two systems (ranging between 2 and 4 THR points) are constant across all the different views of the data. Two exceptions are represented by the `education` and `education equipment` MW terms, for which differences are less marked (respectively 1.33 and 1.8 THR). This seems to indicate that MMT has fewer problems translating the most difficult terms in the data set. At the same time, GT outperforms MMT-I by 5.17 THR in the `possible` MW category, showing that MMT-I probably struggles more than GT for words that might not be terms.

Partial THR			
	GT	MMT-I	MMT-II
All	76.68	74.91	77.23
Disc	80.40	78.83	80.64
Edu	65.33	63.13	67.49
Equip	65.63	63.30	67.13
Sure	77.74	75.94	78.07
Poss	71.27	69.68	72.96

Table 7: Partial THR for GT and the 2 MMT systems. Only Overall scores are reported, since matches are computed at the token level. Results are provided *(i)* for the whole data set (All), *(ii)* by domain category (Disc, Edu, Equip) and *(iii)* for `sure` and `possible` terms.

Comparing MMT results in the two scenarios sheds light on the specific contributions that in-domain data can bring to terminology translation. First of all, in the second scenario there is an increase of the overall performance on the whole data set (+4.36 THR points). The difference with respect to the first scenario is particularly evident for MW terms (+5.75), suggesting that domain adaptation did not only influence lexical choices, but also helped the system to place terms in the correct position. As a matter of fact, if we look at the partial THR results shown in Table 7, we see that the performance gap between the two systems is narrower. This means that the generic and the adapted MMT systems perform similarly in the generation of the SWs composing a MW, but adapted MMT is better at generating them in the correct order. For example, in one of the segments the annotated MW *classification of living beings* was correctly generated in the second scenario, while in the first one the system produced the MW

living classification, which is a match only in the partial THR evaluation.

Finally, the biggest improvement can be found for `education` and `equipment` MW terms, which – as we have seen above – are the most challenging for the MT systems.

As a final observation holding for all systems in both THR evaluations, there is a clear drop in performance when progressing from the evaluation of `sure` terms to that of `possible` terms. The remarkably higher performance obtained on the most reliable terms in the data set highlights the importance of having good quality, flexible gold standards to evaluate translation of terminology.

5 Conclusion and further work

In this contribution we have presented MAGMATic, a gold standard with manually annotated multi-domain terminology. We have described and analised the annotation process and the methods used to check the annotation reliability, and applied the gold standard to the evaluation of NMT in the institutional academic domain.

Given its large size, MAGMATic is able to cover 20 disciplinary domains with a considerable amount of terms each, as well as the education and education equipment domains. Both single-word and multi-word terms are included in this data set, and further distinguished between sure and possible terms. Thanks to these peculiarities, MAGMATic can fill a gap in the field of MT evaluation, providing a valuable test set for insightful and sound quality assessments based on terminology translation. Besides fitting the purpose of evaluating terminology in an MT output, MAGMATic can also be applied to different use cases, e.g. bilingual terminology extraction from word-aligned bilingual corpora where one of the two languages is English, or domain identification in multi-domain English corpora.

The results obtained with adaptive MT on the translation of course catalogues are encouraging, especially taking into account that this is a first attempt to apply NMT to this scenario, and considering the scarcity of available bilingual data. We believe that further work in this field is therefore warranted. From the point of view of MT evaluation, a manual assessment in terms of fluency and adequacy of the outputs produced by MMT and GT could be carried out and its results compared to those described here. This could provide interesting insights into the relationship between correct/incorrect terminology translation and translation quality as perceived by humans. From the point of view of the application scenario, further analyses will be carried out within the second scenario in order to better understand the specific contribution of the in-domain data from each university to the other universities. Finally, in the long-term perspective, we will be able to collect more in-domain data to evaluate the corresponding performance trends of adaptive MT.

References

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

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Astrakhantsev, Nikita A., Denis G. Fedorenko, and Denis Yu. Turdakov. 2015. Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software*, 41(6):336–349.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267. Association for Computational Linguistics.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2018. Neural versus phrase-based MT quality: An in-depth analysis on English-German and English-French. *Computer Speech & Language*, 49:52–70.

Bernier-Colborne, Gabriel and Patrick Drouin. 2014. Creating a test corpus for term extractors through term annotation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 20(1):50–73.

Bertoldi, Nicola, Davide Caroselli, and Marcello Federico. 2018. The ModernMT project. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, Alacant, Spain.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Andy Way, and Panayota Georgakopoulou. 2018. Evaluating MT for massive open online courses. *Machine Translation*, August.

Depraetere, Heidi, Joachim Van den Bogaert, and Joeri Van de Walle. 2011. Bologna translation service: Online translation of course syllabi and study programmes in English. In Forcada, Mikel L., Heidi Depraetere, and Vincent Vandeghinste, editors, *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 29–34, Leuven, Belgium.

Dice, Lee Raymond. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July.

Farajian, M. Amin, Nicola Bertoldi, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Evaluation of terminology translation in instance-based neural MT adaptation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain.

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

Girardi, Christian, Luisa Bentivogli, Mohammad Amin Farajian, and Marcello Federico. 2014. MT-EQuAl: a toolkit for human assessment of machine translation output. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 120–123, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Jalili Sabet, Masoud, Matteo Negri, Marco Turchi, José G.C. de Souza, and Marcello Federico. 2016. TMop: a tool for unsupervised translation memory cleaning. In *Proceedings of ACL-2016 System Demonstrations*, pages 49–54.

Kim, J.-D., T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpusa semantically annotated corpus for biotextmining. *Bioinformatics*, 19(suppl_1):i180–i182, 07.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Papineni, Kishore, 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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Q. Zadeh, Behrang and Siegfried Handschuh. 2014. The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 52–63, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Scott, William A. 1955. Reliability of Content Analysis:The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3):321–325, 01.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Toral, Antonio and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1063–1073.

Van Brussel, Laura, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of NMT, PBMT and RBMT output for english-to-dutch. In Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 3799–3804. European Language Resources Association (ELRA).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Guyon, I., U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.